COORDINATE AND GENERALIZE: A UNIFIED FRAMEWORK FOR AUDIO-VISUAL ZERO-SHOT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-Visual Zero-Shot Learning (AV-ZSL) aims to train a model that can classify videos of unseen classes leveraging audio and visual data, which is achieved by transferring knowledge obtained from seen classes. We identify two imperative issues needed to be addressed: (1) *How to effectively exploit both the audio and visual information?* and (2) *How to transfer the knowledge from seen classes to unseen classes?* In this paper, we ameliorate both of the issues in a unified framework by enhancing two ingredients that existing methods seldom consider. (1) *Multi-Modal Coordination*: Different from existing methods simply fusing the audio and visual features by attention mechanism, we further perform knowledge distillation between the visual and audio branches. This allows information interaction between the two branches and encourages them to learn from each other. (2) *Generalization Capacity*: Existing methods only consider the alignment between the audio-visual features and semantic features on the seen classes, which ignores the generalization capacity. Inspired by the interpretability methods of Deep Neural Networks (DNNs), we propose a novel gradient-based approach to generate transferable masks for the visual and audio features, enforcing the model to focus on the most discriminative segments and benefiting knowledge transfer from seen to unseen classes. Extensive experiments on three challenging benchmarks, ActivityNet-GZSL, UCF-GZSL, and VGGSound-GZSL, demonstrate that our proposed approach can significantly outperform the state-of-the-art methods.

1 INTRODUCTION

With the fast development of Deep Neural Networks (DNNs) (He et al., 2016; Huang et al., 2017), supervised learning has made significant progress over the past few decades. However, its success is primarily attributed to the large amount of labeled data. To this end, Zero-Shot Learning (ZSL) is proposed to classify samples belonging to unseen classes by transferring knowledge obtained from seen classes. Due to its ability to alleviate the reliance of labeled data, ZSL has attracted extensive research attention and been applied in various fields such as image classification (Chen et al., 2017; Bansal et al., 2018; Xian et al., 2018; Bucher et al., 2019; Ramesh et al., 2021), object detection (Bansal et al., 2018; Rahman et al., 2019), semantic segmentation (Bucher et al., 2019; Zhang & Ding, 2021), image generation (Ramesh et al., 2021; Jain et al., 2022) and video classification (Brattoli et al., 2020; Gao et al., 2020).

Most existing zero-shot video classification methods are limited in the visual modality, *i.e.*, only the image sequences of a video are utilized for classification. Considering the natural alignment between the audio and visual data in a video, Audio-Visual Zero-Shot Learning (AV-ZSL) (Parida et al., 2020) is recently introduced to leverage both of the audio and visual information for video classification. Compared to AV-ZSL that only videos belonging to unseen classes appear in the test phase, Audio-Visual Generalized Zero-Shot Learning (AV-GZSL) is a more realistic and challenging variant of AV-ZSL where test videos can come from both the seen and unseen classes. Given the audio and visual data of seen classes available for training, we identify two imperative issues that need to be addressed:

Issue 1: *How to effectively exploit both the audio and visual information?* Existing methods (Mazumder et al., 2021; Mercea et al., 2022b;a) simply fuse the visual and audio features employing cross-modal attention, which is guided by either the cross-entropy loss or the reconstruction loss.

However, the audio and visual information should be complementary and benefit from each other when classifying a video. For example, seeing a piano and hearing the music sound motivates us to classify a video as *PlayingPiano*. Thus, the multi-modal coordination should be considered to effectively exploit both the audio and visual information.

Issue 2: *How to transfer the knowledge obtained from seen classes to unseen classes?* Generalization capacity is at the core of ZSL. Numerous efforts have been made to improve the generalization capacity, *e.g.*, relation modeling (Gao et al., 2019; 2020), uniformity-aware representation learning (Pu et al., 2022), and feature decomposition (Lin et al., 2022; Tong et al., 2019; Li et al., 2021). Existing methods for AV-ZSL, however, only consider the alignment between the audio-visual features and the semantic features. The improvement of the generalization capacity of AV-ZSL models remains to be explored.

In this paper, we tackle both of the issues by (1) introducing mutual knowledge distillation to encourage the multi-modal coordination, (2) proposing a gradient-based method to improve the generalization capacity of the model. Specifically, given audio and visual features extracted from pre-trained models, we propose a two-branch framework to compute similarities with the semantic features of the labels. To encourage the information interaction between the audio and visual branches, mutual knowledge distillation is performed to align the similarity distributions. In addition, inspired by the interpretability methods of DNNs and feature decomposition approaches for ZSL, we elaborate a novel gradient-based approach to improve the generalization capacity. Two transferable masks are generated for the original audio and visual features, which is guided by the gradients relative to the classification loss. This enforces the model to focus on the segment that contributes most to classification, benefiting the knowledge transfer from seen to unseen classes. Our key contributions can be summarized as follows:

- We identify two imperative issues that existing methods seldom considered for AV-ZSL, *i.e.*, multi-modal coordination and generalization capacity, and propose a unified framework to tackle both of them.
- We introduce mutual knowledge distillation between the audio and visual branches to make full use of the multi-modal information.
- We propose a gradient-based mask generation method to ensure the knowledge transfer from seen to unseen classes.
- Extensive experiments on three challenging benchmarks demonstrate the effectiveness of our proposed method.

2 Related Work

Our work is relevant to zero-shot video classification, audio-visual zero-shot learning, knowledge distillation, and interpretability of deep neural networks. We briefly review the relevant work below.

Zero-Shot Video Classification. Zero-Shot Video Classification (ZSVC) (Socher et al., 2013) aims to classify videos of unseen classes by transferring knowledge obtained from seen classes, which has attracted increasing research attention over the last few years (Xu et al., 2015; Gao et al., 2019; Brattoli et al., 2020; Lin et al., 2022; Pu et al., 2022). TS-GCN (Gao et al., 2019) proposes a two-stream Graph Convolutional Network framework that can leverage knowledge graphs to model the relationships between action-attribute, action-action, and attribute-attribute. ResT (Lin et al., 2022) presents an end-to-end cross-modal to associate both visual and semantic spaces and develop a semantic transfer scheme to composite unseen visual prototypes, which alleviates information loss and the hubness problem. AURL (Pu et al., 2022) proposes to learn representation awareness of both alignment and uniformity properties for seen and unseen classes. A supervised contrastive loss is introduced to jointly align visual-semantic features and encourage semantic clusters to distribute uniformly. Despite great success, existing ZSVC works are limited in the visual modality and ignore the audio information which is naturally aligned with visual data.

Audio-Visual Zero-Shot Learning. Recently, Audio-Visual Zero-Shot Learning (AV-ZSL) is proposed to employ both of audio and visual information for video classification. AVGZSLNet (Mazumder et al., 2021) proposes cross-modal decoder and composite triplet to enforce the audio and video embeddings to move closer to the corresponding text embeddings. AVCA (Mercea et al.,

2022b) introduces three novel benchmarks for AV-ZSL and proposes a cross-modal attention model to fuse the audio and visual information. TCAF (Mercea et al., 2022a) is built on AVCA and presents a cross-attention transformer framework to additionally leverage the temporal information. Despite some progress, existing methods only consider the alignment between the audio-visual features and the semantic features. Two imperative issues, *i.e.*, multi-modal coordination and generalization capacity, are rarely considered.

Knowledge Distillation. Knowledge distillation (KD) (Hinton et al., 2015) was proposed to transfer information learned from one model to another, which has been actively studied and broadly applied in various fields, *e.g.*, model compression (Wang et al., 2020; Hou et al., 2020), image classification (Xu et al., 2020a; Cheraghian et al., 2021), object detection (Chen et al., 2017; Zheng et al., 2022), and video understanding (Pan et al., 2020; Tang et al., 2021). Existing KD methods can be roughly divided into two categories: (1) logit mimicking (Hinton et al., 2015; Zhou et al., 2018; Zheng et al., 2022), and (2) feature imitation (Wang et al., 2019; Dai et al., 2021; Guo et al., 2021). While logit mimicking methods supervise the output logit of the student classifier by those of the teacher classifier, feature imitation approaches mimic the intermediate-level features from the hidden layers of the teacher model. In this paper, we perform mutual distillation to facilitate the multi-modal coordination between the audio and visual branches for zero-shot learning.

Interpretability of Deep Neural Networks. DNNs have achieved remarkable success in the past decades. However, the end-to-end learning strategy makes DNN representations a black box, facilitating researchers to delve into the interpretability of models. Gradient-based methods (Selvaraju et al., 2017; Lundberg & Lee, 2017; Chattopadhay et al., 2018) are the mainstream for interpreting DNN. Grad-CAM (Selvaraju et al., 2017) uses the gradients of a target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. Grad-CAM++ (Chattopadhay et al., 2018) introduces pixel-wise weighting of the gradients of the output *w.r.t.* a particular spatial position in the final convolutional feature map. Inspired by the interpretability methods of DNNs, we introduce a gradient-based mask generation method to ensure knowledge transfer from seen to unseen classes.

3 PRELIMINARY

In this section, we formulate the AV-ZSL problem. Let $S = \{(\boldsymbol{a}_i^s, \boldsymbol{v}_i^s, \boldsymbol{w}_i^s, y_i^s)_{i=1}^{N_s} | \boldsymbol{a}_i^s \in \mathcal{A}^s, \boldsymbol{v}_i^s \in \mathcal{V}^s, \boldsymbol{w}_i^s \in \mathcal{W}^s, y_i^s \in \mathcal{Y}^s\}$ represent the seen dataset and $U = \{(\boldsymbol{a}_j^u, \boldsymbol{v}_j^u, \boldsymbol{w}_j^u, y_j^u)_{j=1}^{N_u} | \boldsymbol{a}_i^u \in \mathcal{A}^u, \boldsymbol{v}_j^u \in \mathcal{V}^u, \boldsymbol{w}_j^u \in \mathcal{W}^u, y_j^u \in \mathcal{Y}^u\}$ represent the unseen one, where $\boldsymbol{a}_i^s, \boldsymbol{a}_j^u \in \mathbb{R}^{D_A}$ indicate the D_A-dimensional audio feature in the audio space $\mathcal{A}$ that can be obtained from pre-trained audio model; $\boldsymbol{v}_i^s, \boldsymbol{v}_j^u \in \mathbb{R}^{D_V}$ indicate D_V-dimensional visual feature in the visual space $\mathcal{V}$; $\boldsymbol{w}_i^s, \boldsymbol{w}_j^u \in \mathbb{R}^{D_W}$ indicate D_W-dimensional semantic features (*e.g.*,word vectors) in the semantic space $\mathcal{W}$; $\mathcal{Y}^s = \{y_1^s, \ldots, y_{C_s}^s\}$ and $\mathcal{Y}^u = \{y_1^u, \ldots, y_{C_u}^u\}$ is the label sets of the seen and unseen classes in the label space $\mathcal{Y}$, where C_s and C_u are the number of seen and unseen classes, $\mathcal{Y}^s \cap \mathcal{Y}^u = \emptyset$. In AV-ZSL, the objective is to learn a model $f_{ZSL} : (\mathcal{A}, \mathcal{V}) \rightarrow \mathcal{Y}^u$ that can recognize a testing sample belonging to the unseen classes whose samples are not available during training. AV-GZSL is a more realistic and challenging variant of AV-ZSL where a testing sample can come from both of the seen and unseen classes, *i.e.*, the aim of AV-GZSL is to learn a model $f_{GZSL} : (\mathcal{A}, \mathcal{V}) \rightarrow \mathcal{Y}^s \cup \mathcal{Y}^u$.

4 METHOD

Different from traditional zero-shot video classification where only the visual data is utilized, both the audio and visual data are available in AV-ZSL. Therefore, we identify two imperative issues for AV-ZSL: (1) *How to effectively exploit both the audio and visual information?* and (2) *How to transfer the knowledge obtained from seen classes to unseen classes?* In this paper, we tackle both of the issues in a unified framework, as shown in Figure 1. Given audio and visual features extracted from pre-trained models, two transferable masks are first generated for both of the features. We then fuse the masked features by the attention network and embed them into a common space with the semantic features to obtain the prediction scores. In addition, mutual knowledge distillation is performed to encourage information exchange and multi-modal coordination between the audio and visual branches.

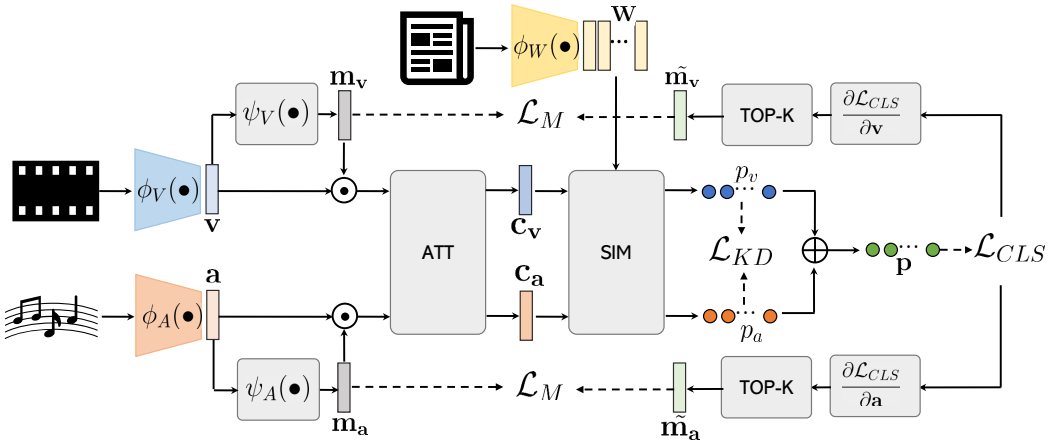

Figure 1: The framework of our proposed method. Given the visual and audio data available for AV-ZSL, we first extract the multi-modal features employing pre-trained models $\phi_V(\bullet)$ and $\phi_A(\bullet)$. Inspired by the interpretability methods, two transferable masks are generated by $\psi_V(\bullet)$ and $\psi_A(\bullet)$ for the visual and audio features, respectively. We then fuse the masked features by an attention network and embed them into a common space with the semantic features to obtain the prediction scores. In addition, mutual knowledge distillation is performed to encourage information interaction and multi-modal coordination between the audio and visual branches. Three loss functions are employed for parameter optimization, namely: the classification loss $\mathcal{L}_{CLS}$, the knowledge distillation loss $\mathcal{L}_{KD}$, and the mask loss $\mathcal{L}_M$.

In the following, we elaborate the main components of our method: (1) mutual knowledge distillation, (2) transferable mask generation, and (3) training and inference.

4.1 Mutual Knowledge Distillation

For a video with visual data V and audio data A, we first extract their visual and audio features using pre-trained models $\phi_V(\bullet)$ and $\phi_A(\bullet)$, *i.e.*, $\boldsymbol{v} = \phi_V(V), \boldsymbol{a} = \phi_A(A)$, where $\boldsymbol{v}$ and $\boldsymbol{a}$ are the visual and audio features, respectively. Then, two transferable masks $\boldsymbol{m}_v$ and $\boldsymbol{m}_a$ are generated by $\psi_V(\bullet)$ and $\psi_A(\bullet)$:

$$\boldsymbol{m}_v = \psi_V(\boldsymbol{v}), \quad \boldsymbol{m}_a = \psi_A(\boldsymbol{a}). \tag{1}$$

Next, we multiply the original features with the generated masks element by element and fuse the multi-modal features using an attention network:

$$\boldsymbol{c}_v, \boldsymbol{c}_a = \mathbf{ATT}((\boldsymbol{v} \odot \boldsymbol{m}_v) \parallel (\boldsymbol{a} \odot \boldsymbol{m}_a)), \tag{2}$$

where $\boldsymbol{c}_v$ and $\boldsymbol{c}_a$ are the learned visual and audio features, respectively. $\mathbf{ATT}$ is the attention network consisting of multi-head attention (Vaswani et al., 2017), feed-forward network, and residual connection. $\odot$ and $\parallel$ represent element-wise multiplication and concatenation, respectively.

Like other zero-shot problems, semantic information is employed to bridge the gap between the seen and unseen classes in AV-ZSL. Specifically, a pre-trained text model $\phi_W(\bullet)$ is utilized to extract the semantic features $\boldsymbol{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{N_s}\}$ from the names of classes, where $\boldsymbol{w}_i$ is the semantic feature of the i-th class and N_s is the number of seen classes. The visual, audio, and semantic features are embedded into a common space to obtain similarity scores:

$$p_v = [p_{v,1}, \ldots, p_{v,N_s}], \quad p_{v,i} = \frac{exp(f_v(\boldsymbol{c}_v) \cdot f_w(\boldsymbol{w}_i)/\tau)}{\sum_k exp(f_v(\boldsymbol{c}_v) \cdot f_w(\boldsymbol{w}_k)/\tau)}, \tag{3}$$

and

$$p_a = [p_{a,1}, \ldots, p_{a,N_s}], \quad p_{a,i} = \frac{exp(f_a(\boldsymbol{c}_a) \cdot f_w(\boldsymbol{w}_i)/\tau)}{\sum_k exp(f_a(\boldsymbol{c}_a) \cdot f_w(\boldsymbol{w}_k)/\tau)}, \tag{4}$$

where p_v and p_a represent the score distributions of visual and audio branches, respectively. $f_v(\bullet)$, $f_a(\bullet)$, and $f_w(\bullet)$ are the embedding networks for the visual, audio, and semantic features, respectively. τ is a temperature parameter to control the sharpness of the distribution.

In order to encourage information interaction between the visual and audio branches, we apply mutual knowledge distillation on the score distribution using the Kullback-Leibler Divergence loss:

$$\mathcal{L}_{KD} = \frac{1}{2}(D_{KL}(p_v \parallel p_a) + D_{KL}(p_a \parallel p_v)), \tag{5}$$

where D_{KL} is the Kullback–Leibler (KL) divergence, *i.e.*, $D_{KL}(p_1 \parallel p_2) = p_1 \log \frac{p_1}{p_2}$. In this way, the visual and audio branches of the model can learn from each other.

4.2 Transferable Mask Generation

After obtaining the p_v and p_a from the visual and audio branches, the final similarity score is obtained by a convex combination:

$$p = \beta p_v + (1 - \beta) p_a. \tag{6}$$

The classification loss is calculated as follows:

$$\mathcal{L}_{CLS} = -\sum_{c=1}^{N_s} (y[c] \log p[c] + (1 - y[c]) \log(1 - p[c])). \tag{7}$$

Inspired by the interpretability of DNNs, we propose a gradient-based mask generation method to facilitate the knowledge transfer from seen to unseen classes. Specifically, the gradients of the classification loss relative to the visual and audio features are computed:

$$\boldsymbol{g}_v = \frac{\partial \mathcal{L}_{CLS}}{\partial \boldsymbol{v}}, \quad \boldsymbol{g}_a = \frac{\partial \mathcal{L}_{CLS}}{\partial \boldsymbol{a}}, \tag{8}$$

where $\boldsymbol{g}_v$ and $\boldsymbol{g}_a$ are the visual and audio gradients, respectively.

To enforce the model focus on the segments that contribute most to the classification, pseudo labels of masks are generated as follows:

$$\tilde{\boldsymbol{m}}_{v,i} = \mathbf{1}_{\boldsymbol{g}_{v,i} \in \mathbb{S}_v}, \quad \tilde{\boldsymbol{m}}_{a,i} = \mathbf{1}_{\boldsymbol{g}_{a,i} \in \mathbb{S}_a}, \tag{9}$$

where $\tilde{\boldsymbol{m}}_v$ and $\tilde{\boldsymbol{m}}_a$ represent the pseudo label of visual and audio masks, respectively. $\mathbb{S}_v$ and $\mathbb{S}_a$ are the sets of top-k elements of the gradients $\boldsymbol{g}_v$ and $\boldsymbol{g}_a$, respectively.

The mask loss is employed to encourage the learned masks to be close with the pseudo ones, which can be formulated as:

$$\mathcal{L}_M = \parallel \tilde{\boldsymbol{m}}_v - \boldsymbol{m}_v \parallel + \parallel \tilde{\boldsymbol{m}}_a - \boldsymbol{m}_a \parallel . \tag{10}$$

4.3 Training and Inference

In summary, three loss functions are employed for parameter optimization, namely: the classification loss $\mathcal{L}_{CLS}$, the knowledge distillation loss $\mathcal{L}_{KD}$, and the mask loss $\mathcal{L}_M$. The overall objective function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{CLS} + \lambda_1 \mathcal{L}_{KD} + \lambda_2 \mathcal{L}_M, \tag{11}$$

where λ_1 and λ_2 are the weighting coefficients for the knowledge distillation loss and the mask loss, respectively. The overall optimization is summarized in Algorithm 1.

During inference, we first obtain the similarity score p and the class prediction c is determined by the maximum value of p:

$$c = \arg\max_j p_j. \tag{12}$$

Algorithm 1 Mutual Distillation and Interpretability Inspired Representation for Audio-Visual Zero-Shot Learning

Input:
Training set $\{(\boldsymbol{a}, \boldsymbol{v}, \boldsymbol{w}, y) | \boldsymbol{a} \in \mathcal{A}^s, \boldsymbol{v} \in \mathcal{V}^s, \boldsymbol{w} \in \mathcal{W}^s, y \in \mathcal{Y}^s\}$, maximum training epochs N_E, batch-size N_B, $\lambda_1 = 1$, and $\lambda_2 = 0.5$.
Output:
Learned model parameters of $\psi_V(\bullet)$, $\psi_A(\bullet)$, $\mathbf{ATT}$, $f_v(\bullet)$, $f_a(\bullet)$, and $f_w(\bullet)$

Initialize parameters of $\psi_V(\bullet)$, $\psi_A(\bullet)$, $\mathbf{ATT}$, $f_v(\bullet)$, $f_a(\bullet)$, and $f_w(\bullet)$
for $t = 1, 2, \ldots, N_E$ **do**
    Sample mini-batch data $\{\boldsymbol{a}_i, \boldsymbol{v}_i, \boldsymbol{w}_i, y_i\}_{i=1}^{N_B}$
    Forward model to generate p_v, p_a, p, $\tilde{\boldsymbol{m}}_v$ and $\tilde{\boldsymbol{m}}_a$
    Calculate $\mathcal{L}_{CLS}$, $\mathcal{L}_{KD}$, and $\mathcal{L}_M$
    $\mathcal{L} \leftarrow \mathcal{L}_{CLS} + \lambda_1 \mathcal{L}_{KD} + \lambda_2 \mathcal{L}_M$
    Jointly update $\psi_V(\bullet)$, $\psi_A(\bullet)$, $\mathbf{ATT}$, $f_v(\bullet)$, $f_a(\bullet)$, and $f_w(\bullet)$ by minimizing $\mathcal{L}$

5 EXPERIMENTS

In this section, all datasets and evaluation protocol are first introduced in detail. Then, we present the implementation details as well as the comparison of experimental results with other state-of-the-art methods. Eventually, ablation study proves the effectiveness of our method.

5.1 DATASETS

We evaluate our method on three recently-introduced benchmarks for AVZSL: ActivityNet-GZSL (Caba Heilbron et al., 2015), UCF-GZSL (Soomro et al., 2012), and VGGSound-GZSL (Chen et al., 2020). The ActivityNet-GZSL dataset covers 200 daily action classes, in which 99, 51, and 50 classes are seen for training, unseen for validation, and unseen for testing, respectively. The UCF-GZSL dataset consists of 30 seen classes, 12 unseen validation classes, and 9 unseen testing classes. The VGGSound-GZSL dataset is a large-scale audio-visual dataset with 138 seen classes, 69 unseen validation classes, and 69 unseen testing classes. The overall statistics of each dataset is reported in Table 1. In order to avoid violating the zero-shot setting, we adopt the dataset splits proposed by AVCA (Mercea et al., 2022b) so that the classes of test samples can be disjoint from those contained in Sports1M (Karpathy et al., 2014) which we used for pre-training the feature extractor.

Table 1: Statistics of the ActivityNet-GZSL, UCF-GZSL, and VGGSound-GZSL datasets.

	ActivityNet-GZSL	UCF-GZSL	VGGSound-GZSL
#Videos	20,168	6,186	93,752
#Classes	200	51	276
#Training Videos	9,204	3,174	70,351
#Seen Validation Videos	1,023	353	7,817
#Unseen Validation Videos	4,307	1,467	3,102
#Seen Testing Videos	1,615	555	9,032
#Unseen Testing Videos	4,199	1,267	3,450
#Seen Classes	99	30	138
#Unseen Validation Classes	51	12	69
#Unseen Testing Classes	50	9	69

5.2 EVALUATION METRICS

In addition to datasets setting, the evaluation metrics are shown as following:

- U : the average per-class accuracy on test videos from unseen classes for AV-GZSL.
- S : the average per-class accuracy on test images from seen classes for AV-GZSL.

Table 2: Comparison with state-of-the-art methods on the ActivityNet-GZSL, UCF-GZSL, and VGGSound-GZSL datasets. $\diamond$ indicates that the temporal features are utilized.

Method	ActivityNet-GZSL				UCF-GZSL				VGGSound-GZSL			
	S	U	H	ZSL	S	U	H	ZSL	S	U	H	ZSL
ALE	2.63	7.87	3.94	7.90	57.59	14.89	23.66	16.32	0.28	5.48	0.53	5.48
SJE	4.61	7.04	5.57	7.08	63.10	16.77	26.50	18.93	**48.33**	1.10	2.15	4.06
DeViSE	3.45	8.53	4.91	8.53	55.59	14.94	23.56	16.09	36.22	1.07	2.08	2.59
APN	9.84	5.76	7.27	6.34	28.46	16.16	20.61	16.44	7.48	3.88	5.11	4.49
f-VAEGAN-D2	4.36	2.14	2.87	2.40	17.29	8.47	11.37	11.11	12.77	0.95	1.77	1.91
CJME	5.55	4.75	5.12	5.84	26.04	8.21	1.48	8.29	8.69	4.78	6.17	5.16
AVGZSLNet	8.93	5.04	6.44	5.40	52.52	10.90	18.05	13.65	18.05	3.48	5.83	5.28
AVCA	24.86	8.02	12.13	9.13	51.53	18.43	27.15	20.01	14.90	4.00	6.31	6.00
TCAF $\diamond$	18.70	7.50	10.71	7.91	58.60	21.74	31.72	24.81	9.64	5.91	7.33	6.06
Ours	**25.13**	**9.46**	**13.75**	**9.62**	**67.41**	**23.88**	**35.27**	**28.22**	20.28	**5.94**	**9.19**	**7.28**

- H : the harmonic mean value for AV-GZSL, which is formulated as:

$$H = \frac{2 \times U \times S}{U + S} \tag{13}$$

- ZSL : the average per-class accuracy on test videos from unseen classes for AV-ZSL.

5.3 Implementation Details

Following the previous method (Mercea et al., 2022b), we utilize a pre-trained SeLaVi (Asano et al., 2020) model to extract the audio and visual features. The dimension of audio and visual features, *i.e.*, D_A and D_V, are set to 512. The semantic features are obtained by word2vec embedding with the size $D_W = 300$. The mask generation networks $\psi_V(\bullet), \psi_A(\bullet)$ and embedding networks $f_v(\bullet), f_a(\bullet), f_w(\bullet)$ are two-layer MLPs. The size of the output of the attention network and the latent space are set to 300 and 64, respectively.

Our approach is implemented with PyTorch (Paszke et al., 2019) and optimized by ADAM (Kingma & Ba, 2014) optimizer with a learning rate of 0.001. The weighting coefficients λ_1 and λ_2 in Equation 11 are determined by the grad search and set to 1 and 0.5, respectively. The proportions of masked locations in the mask generation are set to 0.1, 0.1, and 0.2 for the ActivityNet-GZSL, UCF-GZSL, and VGGSound-GZSL datasets, respectively. The temperature parameter τ in Equation 3,4 and β in Equation 6 are set to 1 and 0.5, respectively. In addition, we set the batch size to 256.

Following previous methods (Mercea et al., 2022b;a), we adopt a two-stage training and evaluation protocol. In the first stage, we train our models on the training set and evaluate on the validation set to determine the hyperparameters. In the second stage, we re-train the models on the union set of training and validation sets using the hyperparameters from the first stage. The final performances are evaluated on the testing set.

5.4 Comparison with State-of-the-Art Methods

Compared Methods. We compare our proposed method with nine approaches adopted for the same task. Among them, ALE (Akata et al., 2015a), SJE (Akata et al., 2015b), DeViSE (Frome et al., 2013), APN(Xu et al., 2020b), and f-VAEGAN-D2 (Xian et al., 2019) are originally proposed for zero-shot image classification. We concatenate the audio and visual features to replace the image features for AV-ZSL. CJME (Parida et al., 2020), AVGZSLNet (Mazumder et al., 2021), AVCA (Mercea et al., 2022b), and TCAF (Mercea et al., 2022a) are the state-of-the-art AV-ZSL approaches.

ALE, SJE, and DeViSE are classical embedding-based ZSL methods, which aim to learn a linear or nonlinear mapping function to measure the compatibility between the input features and label embeddings.

APN proposes an attribute prototype network to jointly learn global and local features. While a visual-semantic embedding layer learns global features, local features are learned through an at-

Table 3: Ablation study of the effectiveness of mutual distillation and interpretability inspired representation on the ActivityNet-GZSL, UCF-GZSL, and VGGSound-GZSL datasets.

Method	ActivityNet-GZSL				UCF-GZSL				VGGSound-GZSL			
	S	U	H	ZSL	S	U	H	ZSL	S	U	H	ZSL
$\mathcal{L}_{CLS}$	23.81	5.06	8.35	5.93	56.09	15.54	24.34	15.87	15.46	3.62	5.87	4.17
$+\mathcal{L}_{KD}$	23.65	7.79	11.72	8.47	66.94	22.49	33.67	23.56	15.70	4.32	6.76	4.81
$+\mathcal{L}_{KD}+\mathcal{L}_{M}$	**25.13**	**9.46**	**13.75**	**9.62**	**67.41**	**23.88**	**35.27**	**28.22**	**20.28**	**5.94**	**9.19**	**7.28**

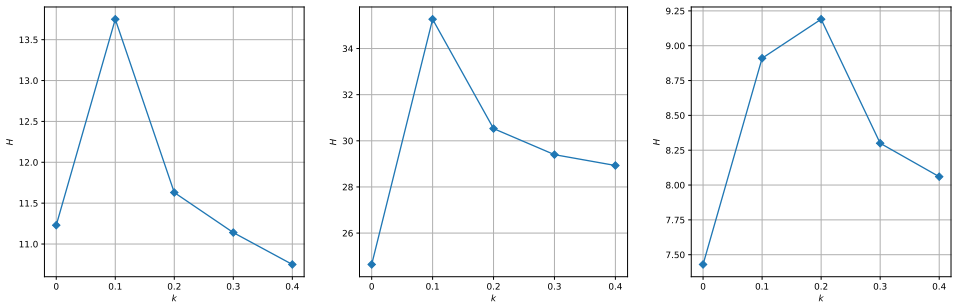

Figure 2: Influence of the masked proportion in mask generation on the ActivityNet-GZSL, UCF-GZSL, and VGGSound-GZSL datasets. We report the harmonic mean value H *w.r.t* the proportion of masked elements.

tribute prototype network that simultaneously regresses and decorrelates attributes from intermediate features.

f-VAEGAN-D2 develops a feature generating framework that synthesizes CNN image features from a class embedding, which circumvents the scarceness of the labeled training data issues and combines conditional VAE and GAN architectures to obtain a more robust generative model.

CJME introduces the task of AV-ZSL and proposes a modality attention-based method to indicate which modality is dominant for classification.

AVGZSLNet proposes cross-modal decoder and composite triplet to enforce the audio and video embeddings to move closer to the corresponding text embeddings.

AVCA proposes a cross-modal attention model to fuse the audio and visual information and align the audio-visual features with the textual label embeddings.

TCAF is built on AVCA and presents a cross-attention transformer framework to additionally leverage the temporal information.

Results. Table 2 shows the results on the ActivityNet-GZSL, UCF-GZSL, and VGGSound-GZSL datasets. It is worth noting that TCAF uses the temporal features while others not, *i.e.*, temporally averaged features are used in other methods and non-averaged ones in TCAF. Our proposed method can outperform previous approaches by a large margin, including TCAF with additional temporal information. For example, compared with previous best method AVCA using the same averaged features, our method increases H and ZSL from 27.15% to 35.27% and from 20.01% to 28.22% respectively on the UCF-GZSL dataset.

5.5 Ablation Study

Effectiveness of mutual distillation and interpretability-inspired representation. In order to evaluate the importance of the proposed mutual distillation and interpretability-inspired representation method, we conduct experiments using different models on the ActivityNet-GZSL, UCF-GZSL, and VGGSound-GZSL datasets. Three variants of our model are trained in this experiment: $\mathcal{L}_{CLS}$, $+\mathcal{L}_{KD}$, and $+\mathcal{L}_{KD}+\mathcal{L}_{M}$. Specifically, $\mathcal{L}_{CLS}$ represents the base model using the classification loss. $+\mathcal{L}_{KD}$ is referred to the base model with the addition of mutual distillation. $+\mathcal{L}_{KD}+\mathcal{L}_{M}$

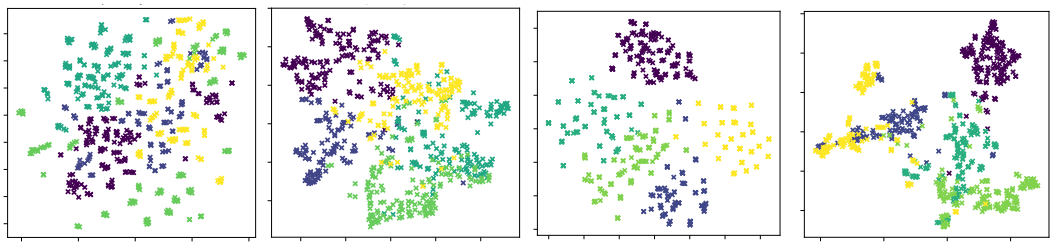

Figure 3: Visualization of the original and learned multi-modal features using t-SNE on the UCF-GZSL dataset. From left to right: the original audio features, the learned audio features, the original visual features and the learned visual features. Different colors represent different unseen classes, *i.e.*, ×: *BandMarching*, ×: *CuttingInKitchen*, ×: *PlayingSitar*, ×: *ShavingBeard*, and ×: *WritingOnBoard*. Best view in color.

is our full model where both of the proposed interpretability-inspired representation and mutual knowledge distillation are utilized.

As shown in Table 3, mutual knowledge distillation can significantly outperform the base model with only the classification loss, which verifies the importance of multi-modal coordination. Interpretability inspired representation can further increase the performance especially ZSL, demonstrating the effectiveness of the mask generation method for improving the generalization capacity.

Influence of the masked proportion in the mask generation. To guide the learning of the transferable masks, pseudo labels are generated by setting the top-k elements of the gradients to 1 and the rest to 0, enforcing the model focus on the segments that contribute most to the classification. In this experiment, we investigate the influence of the masked proportion in mask generation processing. To be specific, we vary the proportion of masked elements from 0 to 0.4 with step 0.1 and report the harmonic mean value H.

Figure 2 shows the results on the ActivityNet-GZSL, UCF-GZSL, and VGGSound-GZSL datasets from left to right. The best performances are obtained when the masked proportions are 0.1, 0.1, and 0.2 for the ActivityNet-GZSL, UCF-GZSL, and VGGSound-GZSL datasets, respectively.

5.6 Qualitative Results

In order to further verify the effectiveness of our proposed method, we visualize the original and learned multi-modal features using t-SNE (Van der Maaten & Hinton, 2008). Specifically, test samples belonging to five of the unseen classes on the UCF-GZSL are employed for visualization: *BandMarching*, *CuttingInKitchen*, *PlayingSitar*, *ShavingBeard*, and *WritingOnBoard*. Figure 3 shows the results of the original audio features, the learned audio features, the original visual features and the learned visual features from left to right. Compared with the original features extracted from pre-trained models, the learned features are better separated and clustered through none samples of unseen classes available during training.

6 Conclusions

In this paper, we focus on the problem of AV-ZSL where both visual and audio data are available for classifying videos belonging to unseen classes. We identify two imperative issues needed to be addressed but ignored by existing methods, *i.e.*, multi-modal coordination and generalization capacity, and propose a mutual distillation and interpretability-inspired representation method. In order to make full use of the multi-modal information, we perform knowledge distillation between the visual and audio branches. Moreover, to facilitate the knowledge transfer from seen to unseen classes, a gradient-based method is elaborated to generate two transferable masks for the visual and audio data, enforcing the model to focus on the segments that contribute most to the classification. Extensive experiments on three benchmarks demonstrate that our proposed approach can significantly outperform the state-of-the-art methods.

REFERENCES

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7): 1425–1438, 2015a.

Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2927–2936, 2015b.

Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems*, 33:4660–4671, 2020.

Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 384–400, 2018.

Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4613–4623, 2020.

Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *NeurIPS2017*, 30, 2017.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *CVPR*, pp. 2534–2543, 2021.

Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *CVPR*, pp. 7842–7851, 2021.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8303–8311, 2019.

Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Learning to model relationships for zero-shot video classification. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3476–3491, 2020.

Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *CVPR*, pp. 2154–2164, 2021.

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*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. Dynabert: Dynamic bert with adaptive width and depth. *NeurIPS*, 33:9782–9793, 2020.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 867–876, June 2022.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Xiangyu Li, Zhe Xu, Kun Wei, and Cheng Deng. Generalized zero-shot learning via disentangled representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1966–1974, 2021.

Chung-Ching Lin, Kevin Lin, Lijuan Wang, Zicheng Liu, and Linjie Li. Cross-modal representation learning for zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19978–19988, June 2022.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *NeurIPS*, 30, 2017.

Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3090–3099, 2021.

Otniel-Bogdan Mercea, Thomas Hummel, A Koepke, and Zeynep Akata. Temporal and cross-modal attention for audio-visual zero-shot learning. *arXiv preprint arXiv:2207.09966*, 2022a.

Otniel-Bogdan Mercea, Lukas Riesch, A. Sophia Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10553–10563, June 2022b.

Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *CVPR*, pp. 10870–10879, 2020.

Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3251–3260, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8026–8037, 2019.

Shi Pu, Kaili Zhao, and Mao Zheng. Alignment-uniformity aware representation learning for zero-shot video classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19968–19977, June 2022.

Shafin Rahman, Salman Khan, and Nick Barnes. Transductive learning for zero-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6082–6091, 2019.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26, 2013.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *NeurIPS*, 34, 2021.

Bin Tong, Chao Wang, Martin Klinkigt, Yoshiyuki Kobayashi, and Yuuichi Nonaka. Hierarchical disentanglement of discriminative latent features for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, pp. 4933–4942, 2019.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *NeurIPS*, 33: 5776–5788, 2020.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10275–10284, 2019.

Kunran Xu, Lai Rui, Yishi Li, and Lin Gu. Feature normalized knowledge distillation for image classification. In *ECCV*, pp. 664–680. Springer, 2020a.

Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33:21969–21980, 2020b.

Xun Xu, Timothy Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 63–67. IEEE, 2015.

Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6974–6983, 2021.

Zhaohui Zheng, Rongguang Ye, Qibin Hou, Dongwei Ren, Ping Wang, Wangmeng Zuo, and Ming-Ming Cheng. Localization distillation for object detection. *arXiv preprint arXiv:2204.05957*, 2022.

Guorui Zhou, Ying Fan, Runpeng Cui, Weijie Bian, Xiaoqiang Zhu, and Kun Gai. Rocket launching: A universal and efficient framework for training well-performing light net. In *AAAI*, volume 32, 2018.