Interspeech 2022
18-22 September 2022, Incheon, Korea

Audio-Visual Generalized Few-Shot Learning with Prototype-Based Co-Adaptation

Yi-Kai Zhang, Da-Wei Zhou, Han-Jia Ye, De-Chuan Zhan

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
{zhangyk, zhoudw, yehj}@lamda.nju.edu.cn, zhandc@nju.edu.cn

Abstract

Although deep learning-based *audio-visual speech recognition* (AVSR) systems recognize base closed-set categories well, extending their discerning ability to additional novel categories with *limited labeled* training data is challenging since the model easily over-fits. In this paper, we propose Prototype-based Co-Adaptation with Transformer (PROTO-CAT), a multi-modal *generalized few-shot learning* (GFSL) method for AVSR systems. In other words, PROTO-CAT learns to recognize a novel class multi-modal object with few-shot training data, while maintaining its ability on those base closed-set categories. The main idea is to transform the prototypes (*i.e.*, class centers) by incorporating cross-modality complementary information and calibrating cross-category semantic differences. In particular, PROTO-CAT co-adapts the embeddings from audio-visual and category levels, so that it generalizes its predictions on all categories dynamically. PROTO-CAT achieves state-of-the-art performance on various AVSR-GFSL benchmarks. The code is available at https://github.com/ZhangYikaii/Proto-CAT.

Index Terms: audio-visual speech recognition, generalized few-shot learning

1. Introduction

McGurk effect [1] demonstrated that visual inputs enormously influence human auditory perception. For example, someone may confuse "bar" and "far" by listening but easily differentiate them given the lip movements videos. Audio-visual speech recognition (AVSR) [2, 3, 4] takes advantage of the rich complementary information from the visual modality for better speech recognition. Many real-world applications include AVSR tasks, like improving speech recognition in multi-talker environments [5], re-dubbing archival silent film [6], serving security systems, and assisting hard-of-hearing people with lipreading devices [7]. Earlier studies exploit HMM with designed handcrafted features [8, 9] to model the temporal dependencies, whereas recently, researchers have adopted deep neural networks and achieved impressive results [10, 11, 12].

Successfully training a deep AVSR model requires manually collecting and labeling massive amounts of data, which incurs immense computational costs. However, in real applications, we expect the model to efficiently recognize novel categories through a few labeled examples [13, 14, 15, 16, 17]. Take intelligent robots as an example [18], where an AVSR model trained on many common words is embedded. Some users may customize the robot with personal instructions like wake-up-words, but only provide a few demonstrations. Thus, the robot should recognize the additional novel vocabularies while maintaining its recognition performance on those old common words. Similar demands also exist in voiceprint recognition [19], buzzwords classification [6], *etc.*, but most deep AVSR models fail in such scenarios since they over-fit on the data-scarce novel categories [20, 21, 22]. Therefore, the *Generalized few-shot learning* (GFSL) [23, 24, 25] ability is essential for an AVSR model.

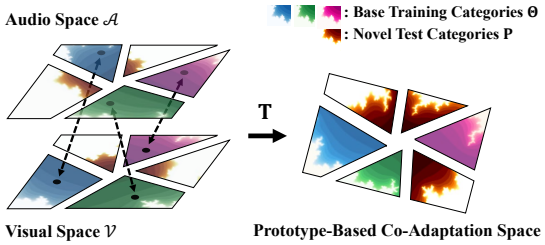


Figure 1: ***Illustration of* PROTO-CAT.** *The model transforms the classification space using* **T** *based on two kinds of audio-visual prototypes (class centers): (1) the base training categories (color with blue, green, and pink); and (2) the additional novel test categories (color with burning).* PROTO-CAT *learns and generalizes on novel test categories from limited labeled examples, maintaining performance on the base training ones.* **T** *includes audio-visual level and category level prototype-based co-adaptation. From left to right, more coverage and more bright colors represent a more reliable classification space.*

In Audio-Visual Generalized Few-Shot Learning (AV-GFSL), we have a base class set with plenty of examples on both visual and audio modalities. We aim to learn a "learning strategy" on the base class set that facilitates the classifier construction on a novel class set even with limited few-shot training data. Besides, the model does not sacrifice any recognition performance of those base ones. The main challenge of AV-GFSL is how to utilize complementary audio-visual information dynamically and calibrate the prediction between the base and novel class sets adaptively.

In this paper, we implement the multi-modal few-shot learning strategy via metric-based classification and learns generalizable embeddings from the base class set. We propose Prototype-based Co-Adaptation with Transformer (PROTO-CAT, see Fig. 1 for illustration), which co-adaptation the prototypes (*i.e.*, class centers) at both *audio-visual level* and *category level*. In detail, the *audio-visual level* transformation selects from audio or visual modality to accurately describes the data. Simultaneously, *category level* adaptation calibrates the predictions between the base and novel categories to enhance the joint classification capability cooperatively. PROTO-CAT and its variants achieve state-of-the-art performance on various AV-GFSL benchmarks.

2. Methods

We first describe the Audio-Visual Generalized Few-Shot Learning (AV-GFSL) task and then discuss how PROTO-CAT

10.21437/Interspeech.2022-652

co-adapts prototypes from the *audio-visual* and *category level*.

2.1. Audio-Visual Generalized Few-Shot Learning

The base class set $\mathcal{B}$ contains examples $\left\{(\boldsymbol{x}_i^{\mathcal{A}}, \boldsymbol{x}_i^{\mathcal{V}}), \boldsymbol{y}_i\right\}$, where $\mathcal{A}$ and $\mathcal{V}$ denotes audio and visual modality. $\boldsymbol{y}_i$ is the one-hot label in $\mathcal{B}$. As mentioned in § 1, AV-GFSL model learns the way to construct an effective classifier given few-shot training data of $\mathcal{B}$ and generalizes its ability to non-overlapping novel class set $\mathcal{N}$. During the test, the AV-GFSL model is provided with N-way K-shot support set with novel classes (K is small, like 1 or 5), *i.e.*,

$$\mathcal{S}_{\text{novel}} = \bigcup_{\boldsymbol{y}_i \in \mathcal{N}} \left\{(\boldsymbol{x}_i^{\mathcal{A}}, \boldsymbol{x}_i^{\mathcal{V}}), \boldsymbol{y}_i\right\}_{i=1}^{K} . \tag{1}$$

There are N classes and K examples per class in $\mathcal{S}_{\text{novel}}$. Then the model is required to discern an M-shot query set with examples from both base and novel classes, *i.e.*,

$$\mathcal{Q}_{\text{novel}} = \bigcup_{\boldsymbol{y}_j \in \mathcal{B} \cup \mathcal{N}} \left\{(\boldsymbol{x}_j^{\mathcal{A}}, \boldsymbol{x}_j^{\mathcal{V}}), \boldsymbol{y}_j\right\}_{j=1}^{M} . \tag{2}$$

AV-GFSL aims to learn a classifier $f\left((\boldsymbol{x}_j^{\mathcal{A}}, \boldsymbol{x}_j^{\mathcal{V}}); \mathcal{S}_{\text{novel}}\right)$ to predict $(\boldsymbol{x}_j^{\mathcal{A}}, \boldsymbol{x}_j^{\mathcal{V}})$ (of either base or novel classes) given the novel class support set $\mathcal{S}_{\text{novel}}$. f would achieve high classification accuracy on $\mathcal{Q}_{\text{novel}}$.

2.2. Prototypical Network and its AV-GFSL Extension

Benefited from meta-learning [26, 13], we simulate the novel class audio-visual few-shot task with examples in the base classes. In detail, we randomly sample the "fake" N-way K-shot support set $\mathcal{S}_{\text{base}} = \bigcup_{\boldsymbol{y}_i \in \mathcal{B}'} \left\{(\boldsymbol{x}_i^{\mathcal{A}}, \boldsymbol{x}_i^{\mathcal{V}}), \boldsymbol{y}_i\right\}_{i=1}^{K}$, where $\mathcal{B}'$ is a size N subset of the base class set $\mathcal{B}$. Compared to the traditional approach of training a whole classifier for the base and novel class set, we find a "learning strategy" to construct the classifier f for each base class few-shot support set and expect it to generalize to the novel class few-shot tasks. Prototypical Network (ProtoNet) [14] is one of the representative *Few-Shot Learning* (FSL) [26, 13, 27, 28, 29, 30] methods, which implements the "learning strategy" as a metric-based classifier *over a single modality*.[1] Given $\mathcal{S}_{\text{base}}$, ProtoNet computes the prototype (*i.e.*, class center) set $\boldsymbol{P}_{\mathcal{S}_{\text{base}}} \in \mathbb{R}^{|\mathcal{B}'| \times d}$ for all classes in $\mathcal{B}'$:

$$\boldsymbol{P}_{\mathcal{S}_{\text{base}}, c} = \frac{1}{K} \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{S}_{\text{base}}} \boldsymbol{\phi}\left(\boldsymbol{x}_i\right) \cdot \mathbb{I}\left[\boldsymbol{y}_i = c\right] . \tag{3}$$

$\boldsymbol{\phi}\left(\cdot\right) : \mathbb{R}^D \to \mathbb{R}^d$ is the embedding function, *i.e.*, feature extractor, and $\mathbb{I}\left[\boldsymbol{y}_i = c\right]$ selects examples of class c. After sampling another M-shot query set, *i.e.*, $\mathcal{Q}_{\text{base}} = \bigcup_{\boldsymbol{y}_j \in \mathcal{B}'} \left\{(\boldsymbol{x}_j^{\mathcal{A}}, \boldsymbol{x}_j^{\mathcal{V}}), \boldsymbol{y}_j\right\}_{j=1}^{M}$, corresponding to $\mathcal{S}_{\text{base}}$, ProtoNet predicts a query example $\boldsymbol{q}_j$ with a distance $\mathbf{d}\left(\cdot, \cdot\right)$:

$$\Pr\left(\hat{\boldsymbol{y}}_j = c \mid \boldsymbol{q}_j; \boldsymbol{P}_{\mathcal{S}_{\text{base}}}\right) = \frac{\exp\left(-\mathbf{d}\left(\boldsymbol{\phi}(\boldsymbol{q}_j), \boldsymbol{P}_{\mathcal{S}_{\text{base}}, c}\right)\right)}{\sum_{c \in \mathcal{B}'} \exp\left(-\mathbf{d}\left(\boldsymbol{\phi}(\boldsymbol{q}_j), \boldsymbol{P}_{\mathcal{S}_{\text{base}}, c}\right)\right)} , \tag{4}$$

so the closer an example to a particular prototype, the larger the probability it belongs to that class. By minimizing the query set classification accuracy over plenty of sampled "fake" tasks, ProtoNet learns discriminative embedding and could be used for a novel class few-shot task $\mathcal{S}_{\text{novel}}$ at test time — computing $\Pr\left(\hat{\boldsymbol{y}}_i \mid \boldsymbol{q}_i; \boldsymbol{P}_{\mathcal{S}_{\text{novel}}}\right)$ based on the novel class prototypes $\boldsymbol{P}_{\mathcal{S}_{\text{novel}}}$.

[1] We use $\boldsymbol{x}_i$ to denote a certain modality without loss of generality.

ProtoNet could be extended to GFSL easily, and we denoted its GFSL version as ProtoNet-GFSL. We compute the prototype set $\boldsymbol{\Theta} \in \mathbb{R}^{|\mathcal{B}| \times d}$ of all base classes. So when we concatenate $\boldsymbol{\Theta}$ with $\boldsymbol{P}_{\mathcal{S}_{\text{novel}}}$, we obtain a generalized classifier for both base and novel classes. The prediction in Eq. 4 is reformulated as $\Pr\left(\hat{\boldsymbol{y}}_i \mid \boldsymbol{q}_i; \boldsymbol{\Theta} \cup \boldsymbol{P}_{\mathcal{S}_{\text{novel}}}\right)$. We can apply ProtoNet-GFSL to each audio and visual modality to handle AV-GFSL tasks. $\boldsymbol{\Theta}$ in ProtoNet-GFSL, however, is fixed for all tasks and imbalanced among classes, which severely limits its ability [23, 24].

2.3. Prototype-based Co-Adaptation with Transformer

Our motivations are: (1) We make the $\boldsymbol{\Theta}$ in Proto-GFSL, as well as the joint GFSL classifier adaptive conditioned on a given few-shot task; (2) We explore complementary information with intra-modal and cross-modal transformations.

The GFSL learning paradigm of PROTO-CAT. As we mentioned before, we can construct a joint classifier for both base and novel classes with their corresponding prototypes. When training PROTO-CAT, we sample an N-way "fake" few-shot task with support set $\mathcal{S}_{\text{base}}$, whose classes are in $\mathcal{B}'$. Given $\mathcal{S}_{\text{base}}$, we construct the prototype set $\boldsymbol{P}_{\mathcal{S}_{\text{base}}}$ of the "fake" novel classes. For the remaining $|\mathcal{B}| - N$ classes, we mask the N classes $\mathcal{B}'$ in $\boldsymbol{\Theta}$ and treat the remaining part of $\boldsymbol{\Theta}$ as the prototypes for "fake" base classes. In summary, we have a joint metric-based classifier $\boldsymbol{\Lambda}$ for both "fake" base and novel classes:

$$\boldsymbol{\Lambda} = \left(\boldsymbol{\Theta} \setminus \bigcup_{c \in \mathcal{B}'} \{\boldsymbol{\Theta}_c\}\right) \cup \boldsymbol{P}_{\mathcal{S}_{\text{base}}} . \tag{5}$$

The prototype $\boldsymbol{\Lambda}$ could be computed for both audio and visual modality, and we construct a cross-modal classifier later based on them. To evaluate the GFSL classifier, we sample another M-shot query set $\mathcal{Q}_{\text{base}}$. It contains examples from both the $|\mathcal{B}| - N$ "fake" base classes and N "fake" novel classes. We repeat the sampling of $\mathcal{S}_{\text{base}}$ and $\mathcal{Q}_{\text{base}}$, *i.e.*, a constructed episode, to create one training epoch.

Co-Adaptation on two levels. In PROTO-CAT, we adapt the prototypes from the *audio-visual level* and *category level*.[2]

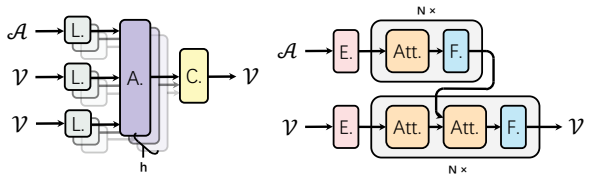
At the *category level*, PROTO-CAT adapts the joint metric-based classifier for both base and novel classes with co-adaptation on $\boldsymbol{\Lambda}$. Based on Transformer [31], PROTO-CAT introduces a modified Multi-Head Attention module to calibrate the two sets of prototypes. Concretely, we treat $\boldsymbol{\Lambda}$ as query, key, and value in Transformer to calibrate the prototypes. For one modality, audio or visual, the intra-modal transformation $\mathbf{T}_{\text{intra}}$ is defined as:

$$\begin{aligned} \mathbf{T}_{\text{intra}}\left(\boldsymbol{\Lambda}\right) &= \boldsymbol{\Lambda} + \boldsymbol{\alpha}\left(\mathtt{Q},\mathtt{K},\mathtt{V}=\boldsymbol{\Lambda}\right) \\ &= \boldsymbol{\Lambda} + \operatorname{softmax}\left(\frac{\boldsymbol{\Lambda}\mathrm{W}^{\mathtt{Q}} \cdot \left(\boldsymbol{\Lambda}\mathrm{W}^{\mathtt{K}}\right)^{\top}}{\sqrt{d}}\right)\boldsymbol{\Lambda}\mathrm{W}^{\mathtt{V}} . \end{aligned} \tag{6}$$

In Eq. 6, we apply linear projections on the query, key, and values using $\mathrm{W}^{\mathtt{Q}}$, $\mathrm{W}^{\mathtt{K}}$, and $\mathrm{W}^{\mathtt{V}}$, respectively. The similarity between prototypes is measured by inner product in the transformed space, which results in larger weights of the attention head $\boldsymbol{\alpha}$. Here d is the size of every attention head.

To make better use of complementary modality information, PROTO-CAT also constructs the *audio-visual level* adaptation with cross-modal transformation $\mathbf{T}_{\text{cross}}$. It adapts the audio

[2] We omit the modality notation when introducing a particular modality, and use the superscript to emphasize the interaction between modalities later.



(a) Cross-Att. of PROTO-CAT (b) Cross-Att. of PROTO-CAT$^+$

Figure 2: ***Overall Architecture of Cross-Modal Attention***. *Here $\mathcal{A}$ and $\mathcal{V}$ represent the audio and visual modality, respectively.* PROTO-CAT *(left) construct Multi-Head Attention with* h *heads with Linear (*L.*), Scaled Dot-Product Attention (*A.*), and Concat (*C.*) modules.* PROTO-CAT$^+$ *(right) is Transformer-based with* N *stacked Category Insertion Encoding (*E.*), Multi-Head Attention (*Att.*), and Feed Forward (*F.*) followed by Normalization layers.*

embeddings to the same semantics visual ones towards learning modality relevance. As shown in Fig. 2(a), $\mathbf{T}_{\text{cross}}$ includes cross-modal key-query-value pairs:

$$\mathbf{T}_{\text{cross}}\left(\mathbf{\Lambda}^{\mathcal{A}}, \mathbf{\Lambda}^{\mathcal{V}}\right) = \mathbf{\Lambda}^{\mathcal{A}} + \boldsymbol{\alpha}\left(\texttt{Q=}\mathbf{\Lambda}^{\mathcal{A}}, \texttt{K,V=}\mathbf{\Lambda}^{\mathcal{V}}\right) . \quad (7)$$

We also consider an improved model PROTO-CAT$^+$ using Transformer-based Modal Translator [32] (as illustrated in Fig. 2(b)). PROTO-CAT$^+$ translates the audio embeddings to visual ones with encoder-decoder architecture, *i.e.*, the audio and visual part is inputted as the source and target sequence, respectively. It highlights the ability to capture deep cross-modal information in complex environments.

Training objective for PROTO-CAT. We have three types of transformed prototypes: the cross-modal one and the other two transformed by intra-modal attention. We denote them as a set $\boldsymbol{\mu} = \left\{\mathbf{T}_{\text{intra}}\left(\mathbf{\Lambda}^{\mathcal{A}}\right), \mathbf{T}_{\text{intra}}\left(\mathbf{\Lambda}^{\mathcal{V}}\right), \mathbf{T}_{\text{cross}}\left(\mathbf{\Lambda}^{\mathcal{A}}, \mathbf{\Lambda}^{\mathcal{V}}\right)\right\}$. Our models classify query example $\boldsymbol{q}_i$ to class c by fusing nearest neighbor probabilities based on Eq. 4:

$$\Pr\left(\hat{\boldsymbol{y}}_i = c \mid \boldsymbol{q}_i; \boldsymbol{\mu}\right) = \frac{1}{3}\sum_{\boldsymbol{\mu}_k \in \boldsymbol{\mu}} \Pr\left(\hat{\boldsymbol{y}}_i = c \mid \boldsymbol{q}_i; \boldsymbol{\mu}_k\right) . \quad (8)$$

Alg. 1 shows the overall training process for each episode. Classifier $f\left(\cdot\,; \mathbf{\Theta}, \mathcal{S}_{\text{base}}\right)$ is summarized from the above process. We optimize it by minimizing the average error over the prediction on $\mathcal{Q}_{\text{base}}$:

$$f^* = \arg\min_{f} \sum_{\left\{\left(\boldsymbol{x}_i^{\mathcal{A}}, \boldsymbol{x}_i^{\mathcal{V}}\right), \boldsymbol{y}_i\right\} \in \mathcal{Q}_{\text{base}}} \ell\Big(f\left(\left(\boldsymbol{x}_i^{\mathcal{A}}, \boldsymbol{x}_i^{\mathcal{V}}\right); \mathbf{\Theta}, \mathcal{S}_{\text{base}}\right), \boldsymbol{y}_i\Big) , \quad (9)$$

where $\ell\left(\cdot\,,\,\cdot\right)$ is the loss function. f is the classifier of PROTO-CAT with the parameters $\mathbf{\Theta}$ and given support set $\mathcal{S}_{\text{base}}$. At test time, the optimized classifier $f^*\left(\cdot\,; \mathbf{\Theta}, \mathcal{S}_{\text{novel}}\right)$ will calculate $\mathbf{\Lambda}$ as $\mathbf{\Theta} \cup \boldsymbol{P}_{\text{novel}}$ (The latter comes from $\mathcal{S}_{\text{novel}}$), and perform a series of transformations as Eq. 6, 7. Finally, it predicts on the novel class query set $\mathcal{Q}_{\text{novel}}$ w.r.t. Eq. 8.

Algorithm 1 Model Training for PROTO-CAT

Require: Training data on the base class set $\mathcal{B}$.
for each episode of epoch **do**
  Sample task $(\mathcal{S}_{\text{base}}, \mathcal{Q}_{\text{base}})$ as mentioned in § 2.2.
  Compute *adaptive task-specific* $\boldsymbol{P}$ of $\mathcal{S}_{\text{base}}$ as Eq. 3
  $\mathbf{\Lambda} \leftarrow$ Insert $\boldsymbol{P}$ into *neural task-shared* $\mathbf{\Theta}$ as Eq. 5
  Co-Adaptation with Transformer as Eq. 6, 7
    $\boldsymbol{\mu} \leftarrow \left\{\mathbf{T}_{\text{intra}}\left(\mathbf{\Lambda}^{\mathcal{A}}\right), \mathbf{T}_{\text{intra}}\left(\mathbf{\Lambda}^{\mathcal{V}}\right), \mathbf{T}_{\text{cross}}\left(\mathbf{\Lambda}^{\mathcal{A}}, \mathbf{\Lambda}^{\mathcal{V}}\right)\right\}$
  for all query examples $\boldsymbol{q}_i \in \mathcal{Q}_{\text{base}}$ **do**
    Classify $\hat{\boldsymbol{y}}_{i,\text{test}}$ with $\Pr\left(\hat{\boldsymbol{y}}_i \mid \boldsymbol{q}_i; \boldsymbol{\mu}\right)$ as Eq. 8
    Compute $\ell\left(\hat{\boldsymbol{y}}_{i,\text{test}}, \boldsymbol{y}_i\right)$ with Eq. 9
  end for
  Compute $\nabla_{\boldsymbol{\phi},\mathbf{T},\mathbf{\Theta}} \sum_{\boldsymbol{q}_i \in \mathcal{Q}_{\text{base}}} \ell\Big(f\left(\boldsymbol{q}_i; \mathbf{\Theta}, \mathcal{S}_{\text{base}}\right), \boldsymbol{y}_i\Big)$
  Update $\boldsymbol{\phi}, \mathbf{T}, \mathbf{\Theta}$ with $\nabla_{\boldsymbol{\phi},\mathbf{T},\mathbf{\Theta}}$
end for

3. Experiments

3.1. Experimental Setups

Datasets. We study on two public large-scale audio-visual datasets: Lip Reading in the Wild (LRW) [34] and CAS-VSR-W1k (LRW-1000) [35]. LRW and LRW-1000 contain human speaking videos for word-level AVSR in English and Mandarin, respectively. LRW contains 500 different word classes from over 1,000 speakers. LRW-1000 contains 1,000 word classes from more than 2,000 speakers. These datasets cover the variations in scale, resolution, and background clutter to incorporate challenges in a real-world applications.

GFSL Simulation. Neither LRW nor LRW-1000 are naturally generalized few-shot datasets. As mentioned in § 2.1, we randomly split the whole class set as base (for training), novel (for validation), and novel (for testing) with the number of 64, 16 and 20. We keep novel classes out of the training and sample 5-way 1-shot tasks to simulate the GFSL scenario as shown in Eq. 1, 2. Half of the query sets are from the base classes, and half are from the novel ones during testing.

Implementation Details. For data preprocessing, each video sequence is cropped as a fixed 96×96 pixels wide ROI so that the mouth region is roughly centered (LRW-1000 has been pre-cropped). Each audio clip is downsampled to 16 kHz and normalized. All data is firstly forwarded into the feature extraction backbone. The frontend is a modified ResNet-18: for the visual part, its first layer is replaced by a 3D convolutional layer with kernel size $5 \times 7 \times 7$, and the audio one is a 1D convolutional layer. The backend is the LSTM-based [36], GRU-based [37, 38], or Multi-Scale Temporal Convolutional Network (MS-TCN)-based [39].

3.2. Results

Comparison with other benchmarks. As shown in Table 1, we present the comparison between (1) traditional audio-visual embeddings based on LSTM, GRU, and MS-TCN backbones; (2) GFSL extension of gradient-based meta learning models, *e.g.*, MAML [13] and BootstrappedMAML [30]; (3) GFSL extension of metric-based FSL models, *e.g.*, ProtoNet [14], MatchingNet [26], MetaOptNet [27], DeepEMD [28], and FEAT [29]; (4) standard generalized few-shot approaches such as DFSL [23] and CASTLE [24]. For audio-visual recognition results in Table 1, rows 1–3 verifies that embeddings trained with the traditional way tend to fail in predictions on novel classes. Rows 4–5 show that MAML-based linear classifier, as the partially updated global prototypes, slightly improve the novel classes' performance, but at the same time severely loses the base ones. Rows 6 – 10 indicate that the base prototypes computed directly from the entire training set interfere with the classification of novel classes, even though they use well-designed local metric-based classifiers within one task. Rows 11 – 12 mean DFSL and CASTLE alleviate this problem with learning shared weights on base classes. Our methods PROTO-CAT and PROTO-CAT$^+$ significantly outperform the above

Table 1: ***AV-GFSL classification performance*** *(in %; measured over* $10,000$ *rounds; higher is better) of* 5*-way* 1*-shot training tasks on* LRW *and* LRW-1000 *datasets. The best result of each scenario is in bold font. The performance measure on both base and novel classes (Base, Novel) is mean accuracy. Harmonic mean (*H*-mean) of the above two is a better GFSL performance measure [33, 24].*

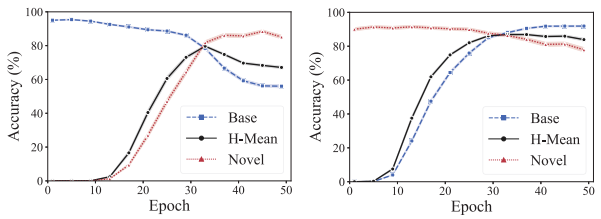
Dataset	**LRW [34]**					**LRW-1000 [35]**		
Data Source	Audio($\mathcal{A}$)	Visual($\mathcal{V}$)	Audio-Visual ($\mathcal{A}\&\mathcal{V}$)			Audio-Visual ($\mathcal{A}\&\mathcal{V}$)		
Perf. Measures on	H-Mean	H-Mean	Base	Novel	H-Mean	Base	Novel	H-Mean
LSTM-based [36]	32.20	8.00	97.09	23.76	37.22	71.34	0.03	0.07
GRU-based [37, 38]	37.01	10.58	**97.44**	27.35	41.71	71.34	0.05	0.09
MS-TCN-based [39]	62.29	19.06	80.96	51.28	61.76	71.55	0.33	0.63
MAML [13]	35.49	10.25	40.09	66.70	49.20	29.40	23.21	25.83
BootstrappedMAML [30]	33.75	6.52	35.29	64.20	45.17	28.15	27.98	28.09
ProtoNet [14]	39.95	14.40	96.33	39.23	54.79	69.33	0.76	1.47
MatchingNet [26]	36.76	12.09	94.54	36.57	52.31	68.42	0.95	1.89
MetaOptNet [27]	43.81	19.59	88.20	47.06	60.73	69.01	1.79	3.44
DeepEMD [28]	–	27.02	82.53	16.43	27.02	64.54	0.80	1.56
FEAT [29]	49.90	25.75	96.26	54.52	68.83	**71.69**	2.62	4.89
DFSL [23]	72.13	42.56	66.10	84.62	73.81	31.68	**68.72**	42.56
CASTLE [24]	75.48	34.68	73.50	90.20	80.74	11.13	54.07	17.84
PROTO-CAT (Ours)	**84.18**	**74.55**	93.37	**91.20**	**92.13**	49.70	38.27	42.25
PROTO-CAT^{+} (Ours)			93.18	90.16	91.49	54.55	38.16	**43.88**

Table 2: ***Ablation studies*** *on 1) audio-visual level intra-modal* $\mathbf{T}_{intra}$ *and cross-modal* $\mathbf{T}_{cross}$ *transformation, and 2) category level transformation operating on different class set parts. On the basis of "Identity Transform", "*$+\ \mathbf{T}_{intra}$*" means only implement intra-modal transformation, "*$+\ \mathbf{T}$ *on Base" means only operate on base class set, and so on. We measure the performance over* $10,000$ *rounds on* LRW *dataset as in Table 1.*

Data Source	Audio-Visual ($\mathcal{A}\&\mathcal{V}$)		
Perf. Measures on	Base	Novel	H-Mean
Audio-Visual Level			
Identity Transform	92.60	44.51	59.53
$+\ \mathbf{T}_{\text{intra}}$	91.79	89.55	90.42
$+\ \mathbf{T}_{\text{intra}}, \mathbf{T}_{\text{cross}}$	**93.37**	**91.20**	**92.13**
Category Level			
Identity Transform	92.60	44.51	59.53
$+\ \mathbf{T}$ on Novel	67.55	83.41	74.36
$+\ \mathbf{T}$ on Base	86.58	72.68	78.85
$+\ \mathbf{T}$ on Both	**93.37**	**91.20**	**92.13**

in terms of the combined harmonic mean of base and novel class sets' accuracy. They achieve AV-GFSL classification performance of up to 92.13% on LRW and 43.88% on LRW-1000. For the cross-modal complementary strengths (audio-video level co-adaptation influence), previous GFSL methods like DFSL (rows 11), with H-mean from 72.13% to 73.81% ($1.68\%\uparrow$), shows the weak growth of joint modality strategy. Our approaches, shown in rows 13 – 14, improve single-modal performance from 84.18% to 92.13% ($7.95\%\uparrow$) and achieve more than 90% accuracy on base and novel mixed classes.

Ablation Study. We further perform the ablation analysis of the *audio-visual* and *category level* prototype-based co-adaptation in Table 2 and Fig. 3(a), 3(b). (1) For *audio-visual level* adaptation with cross-modal transformation component $\mathbf{T}_{\text{cross}}$, we use only $\mathbf{T}_{\text{intra}}$, *i.e.*, "$+\mathbf{T}_{\text{intra}}$" in the upper part of Table 2, whose harmonic mean accuracy drops by 1.71% compared to "$+\mathbf{T}_{\text{intra}}, \mathbf{T}_{\text{cross}}$". Compared to direct calculation ("Identity Transform"), $\mathbf{T}_{\text{intra}}$ and $\mathbf{T}_{\text{cross}}$ significantly improve the performance on the novel class set. (2) We then study the novel-only, base-only, and whole-class set transformation. Fig. 3(a) and Fig. 3(b) show that the transformation $\mathbf{T}$ operates on different parts of the class set, novel or base. Our model adapts to this part continuously and maintains the performance of the other part. The lower part of Table 2 shows the novel-only and base-only results from a quantitative perspective, just as "$\mathbf{T}$ on Novel" performs better on the novel class set part (83.41% is 10.73% more than 72.68%), while "$\mathbf{T}$ on Base" does the opposite. Our approach with "$\mathbf{T}$ on Both" considers adaptation on the entire class set to balance at the category level and achieve optimal performance.



(a) $\mathbf{T}$ only operates on Novel (b) $\mathbf{T}$ only operates on Base

Figure 3: ***Ablation studies*** *of the performance on* LRW *shows transformation* $\mathbf{T}$ *makes the currently operated class set work.*

4. Conclusions

In this paper, we observe that Audio-Visual Generalized Few-Shot Learning (AV-GFSL) provides a more real-world application solution. It requires learning and generalizing on novel test categories while maintaining the discriminative ability on base closed-set categories. We propose to do multi-modal co-adaptation based on the classification prototypes (PROTO-CAT). PROTO-CAT captures the cross-modality complementary information and incorporates it into the cross-category co-adaptation. The results and ablation study highlight PROTO-CAT co-adapts to audio-visual prototypes with strong base-novel classification performance. Our future directions include improving the robustness of the model and developing toward more realistic applications.

Acknowledgments This research was supported by National Key R&D Program of China (2020AAA0109401), NSFC (61773198, 61921006, 62006112), NSFC-NRF Joint Research Project under Grant 61861146001, Collaborative Innovation Center of Novel Software Technology and Industrialization, NSF of Jiangsu Province (BK20200313), CCF-Hikvision Open Fund (20210005).

5. References

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[2] S. P. Panda, "Automated speech recognition system in advancement of human-computer interaction," in *ICCMC*, 2017, pp. 302–306.

[3] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Appl. Intell.*, vol. 42, no. 4, pp. 722–737, 2015.

[4] A. Gupta, Y. Miao, L. Neves, and F. Metze, "Visual features for context-aware speech recognition," in *IEEE ICASSP*, 2017, pp. 5020–5024.

[5] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

[6] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE TPAMI*, 2018.

[7] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," *CoRR*, vol. abs/1611.01599, 2016.

[8] G. I. Chiou and J. Hwang, "Lipreading from color video," *IEEE TIP*, vol. 6, no. 8, pp. 1192–1195, 1997.

[9] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multim.*, vol. 2, no. 3, pp. 141–151, 2000.

[10] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," in *INTERSPEECH*, F. Lacerda, Ed., 2017, pp. 3652–3656.

[11] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *IEEE ICASSP*, 2021, pp. 7613–7617.

[12] S. Ren, Y. Du, J. Lv, G. Han, and S. He, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *IEEE CVPR*, 2021, pp. 13 325–13 333.

[13] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, vol. 70, 2017, pp. 1126–1135.

[14] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, 2017, pp. 4077–4087.

[15] H. Ye, X. Li, and D. Zhan, "Task cooperation for semi-supervised few-shot learning," in *AAAI*, 2021, pp. 10 682–10 690.

[16] D.-W. Zhou, F.-Y. Wang, H.-J. Ye, L. Ma, S. Pu, and D.-C. Zhan, "Forward compatible few-shot class-incremental learning," in *CVPR*, 2022, pp. 9046–9056.

[17] H.-J. Ye, L. Han, and D.-C. Zhan, "Revisiting unsupervised meta-learning via the characteristics of few-shot tasks," *IEEE TPAMI*, 2022.

[18] Y. Wang, N. J. Bryan, M. Cartwright, J. P. Bello, and J. Salamon, "Few-shot continual learning for audio classification," in *IEEE ICASSP*, 2021, pp. 321–325.

[19] L. G. KERSTA, "Voiceprint identification," *Nature*, vol. 196, pp. 1253–1257, 1962.

[20] S. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *IEEE CVPR*, 2017, pp. 5533–5542.

[21] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Co-transport for class-incremental learning," in *ACM MM*, 2021, pp. 1645–1654.

[22] D.-W. Zhou, F.-Y. Wang, H.-J. Ye, and D.-C. Zhan, "Pycil: A python toolbox for class-incremental learning," *arXiv preprint arXiv:2112.12533*, 2021.

[23] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *IEEE CVPR*, 2018, pp. 4367–4375.

[24] H. Ye, H. Hu, and D. Zhan, "Learning adaptive classifiers synthesis for generalized few-shot learning," *IJCV*, vol. 129, no. 6, pp. 1930–1953, 2021.

[25] X. Shi, L. Salewski, M. Schiegg, and M. Welling, "Relational generalized few-shot learning," in *BMVC*, 2020.

[26] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *NIPS*, 2016, pp. 3630–3638.

[27] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *IEEE CVPR*, 2019, pp. 10 657–10 665.

[28] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *IEEE CVPR*, 2020, pp. 12 200–12 210.

[29] H. Ye, H. Hu, D. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *IEEE CVPR*, 2020, pp. 8805–8814.

[30] S. Flennerhag, Y. Schroecker, T. Zahavy, H. van Hasselt, D. Silver, and S. Singh, "Bootstrapped meta-learning," *CoRR*, vol. abs/2109.04504, 2021.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[32] W. Li, D. Jiang, W. Zou, and X. Li, "TMT: A transformer-based modal translator for improving multimodal sequence representations in audio visual scene-aware dialog," in *INTERSPEECH*, 2020, pp. 3501–3505.

[33] E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero- and few-shot learning via aligned variational autoencoders," in *IEEE CVPR*, 2019, pp. 8247–8255.

[34] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *ACCV*, 2016.

[35] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *IEEE FG*, 2019, pp. 1–8.

[36] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *IEEE ICASSP*, 2016, pp. 6115–6119.

[37] J. Xiao, S. Yang, Y. Zhang, S. Shan, and X. Chen, "Deformation flow based two-stream network for lip reading," in *IEEE FG*, 2020, pp. 364–370.

[38] D. Feng, S. Yang, S. Shan, and X. Chen, "Learn an effective lip reading model without pains," *CoRR*, vol. abs/2011.07557, 2020.

[39] B. Martínez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *IEEE ICASSP*, 2020, pp. 6319–6323.