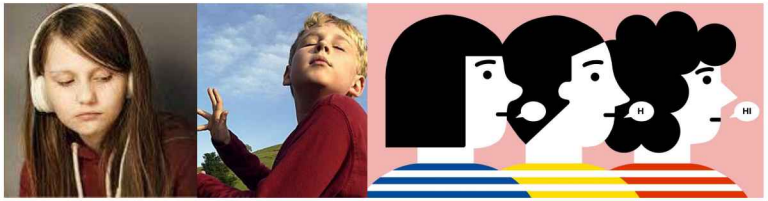
Discriminative Few Shot Learning of Facial Dynamics in ADOS Videos

Na Zhang

- People with Autism Spectrum Disorder (ASD) show pervasive dysfunctions in social and communicative behaviors:
  - eye-contact
  - hand movements
  - speech traits
  - reciprocal social exchange
  - …



- The possible connections between ASD and facial characteristics have been studied
  - Difficulty in interpreting and regulating emotions

Literature

- Most existing works that focus on face analysis are based on images or short videos.
- Few works aim at Autism Diagnostic Observation Schedule (ADOS) videos, due to
  - complexity
    - interaction between participant and examiner
  - length
    - usually last for hours
- In this work, we attempt to fill this gap
  - **facial dynamics feature**

Autism Diagnostic Observation Schedule: ADOS Videos

- Structured but natural discussion
- Different scenes are designed for the analysis of different aspects of autism signs/symptoms

Scene 1

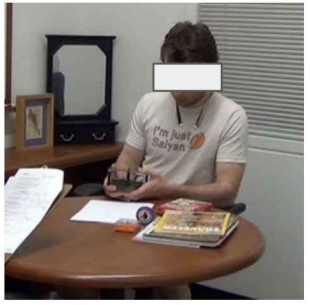
Scene 2

Scene 3-4

Scene 5-7, 11-14

Scene 8

Scene 9



Scene 10

Scene 15

1. Construction Task
2. Telling A Story
3-4. Describing A Picture & Talking
5-7. Conversation on School, Work, Social Difficulties & Emotions
8. Demonstration Task
9. Cartoons
10. Break
11-14. Conversation on Daily Living, Relationships, Plans
15. Creating A Story

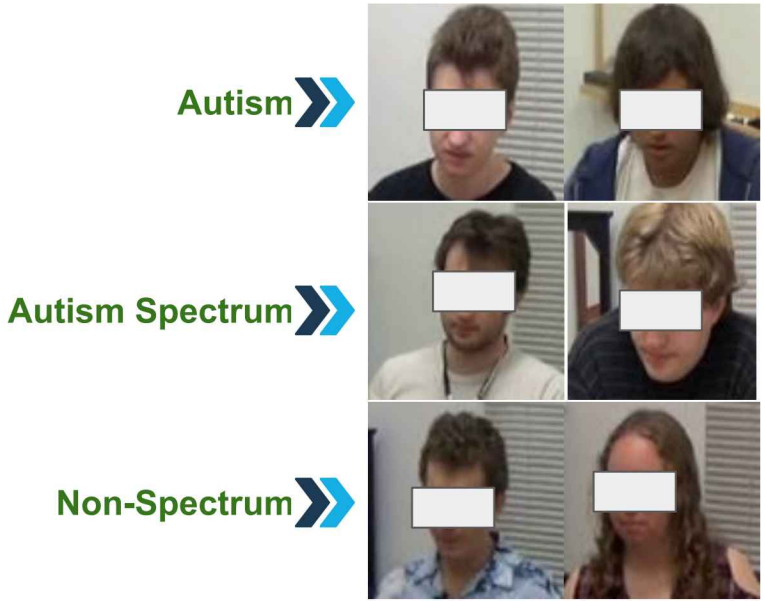
- Each video ranges in 50 ~ 170 minutes
- Captures the complicated and rich behaviors
  - Body behaviors – scene 8, 9
  - face expressions – scene 5-7, 11-14
  - hand gestures – scene 1, 10, 15
  - eye contact – except 1,10
  - speech traits (volume, pacing) – all
  - reciprocal social exchange with the examiner – scene 2-7, 11-14
- We can use different scenes for the analysis of different aspects
  - 5-7, 11-14 high quality faces

Our Selection

1. Construction Task
2. Telling A Story
3-4. Describing A Picture & Talking
5-7. Conversation on School, Work, Social Difficulties & Emotions
8. Demonstration Task
9. Cartoons
10. Break
11-14. Conversation on Daily Living, Relationships, Plans
15. Creating A Story

Labelling & Categories

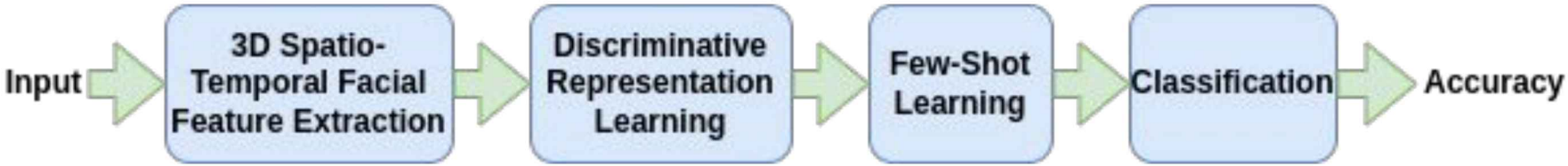
- Detailed scoring is provided by the examiner
- Overall ratings are made by reliable experts
- **ASD: 42 videos**
  - Severity Levels
    - Autism: 17 videos
    - Autism Spectrum: 10 videos
    - Non-Spectrum: 15 videos

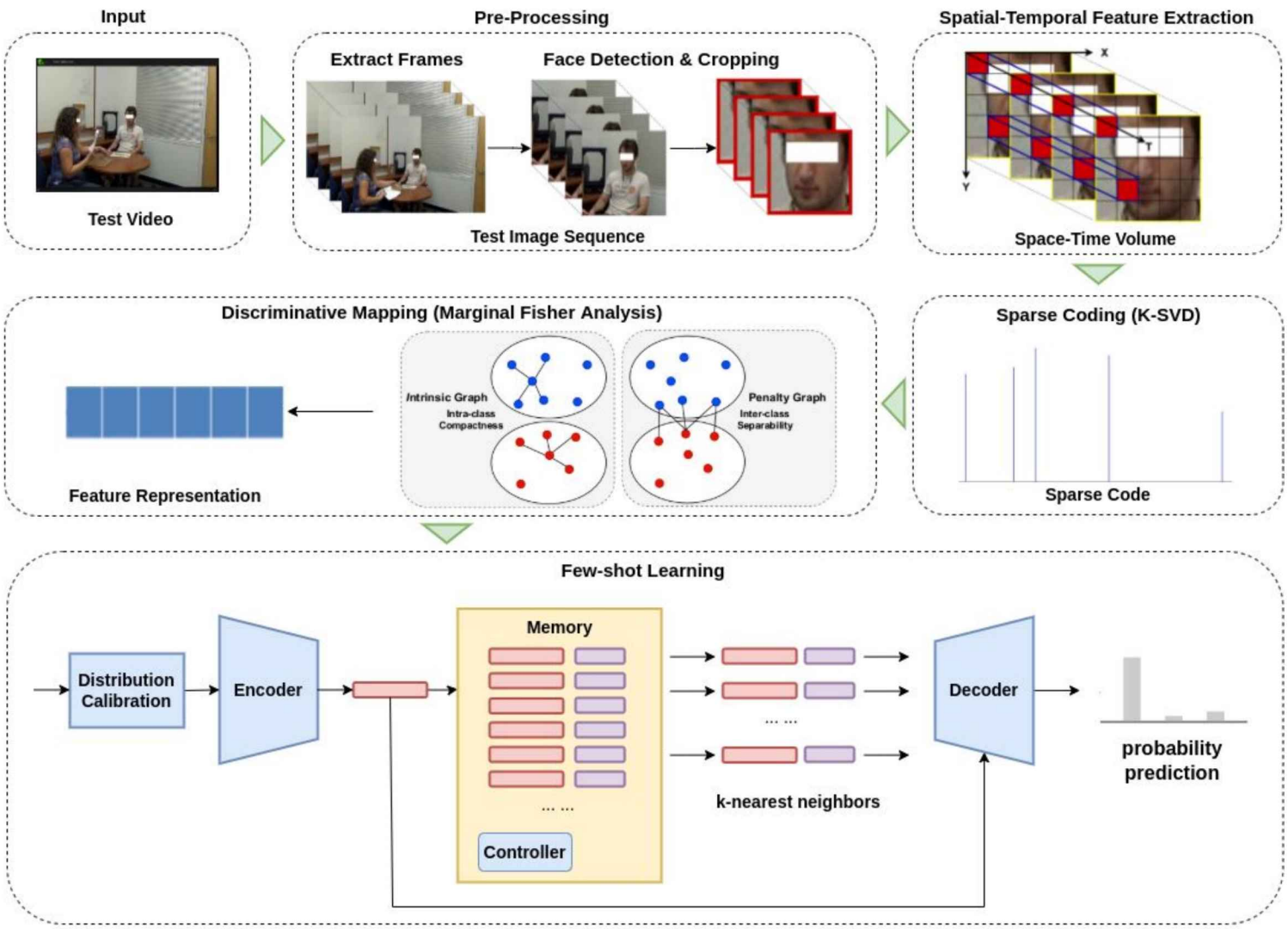


Method

- Develop a discriminative FSL method to capture facial dynamics in data for severity level prediction of autism
  - Spatial
    - facial appearance, static expression, eye movements, etc.
  - Temporal
    - expression changes, gaze patterns, and head pose variations

➢ **Autism (10)** ➢ **Autism Spectrum (17)** ➢ **Non-Spectrum (15)**





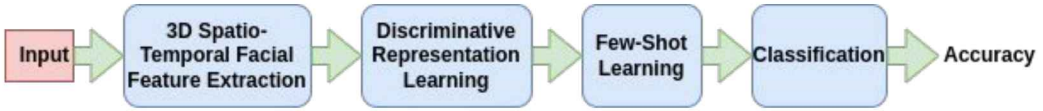
Pre-processing

3D Facial Feature Extraction

Feature Representation learning

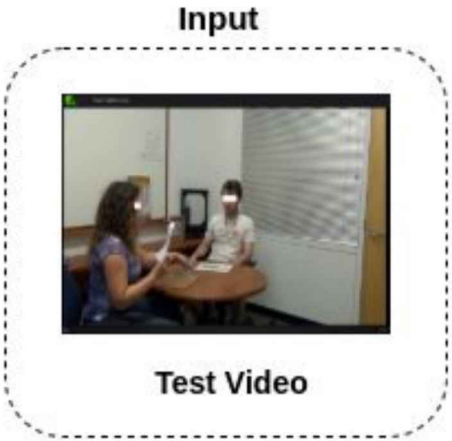
Few Shot Learning

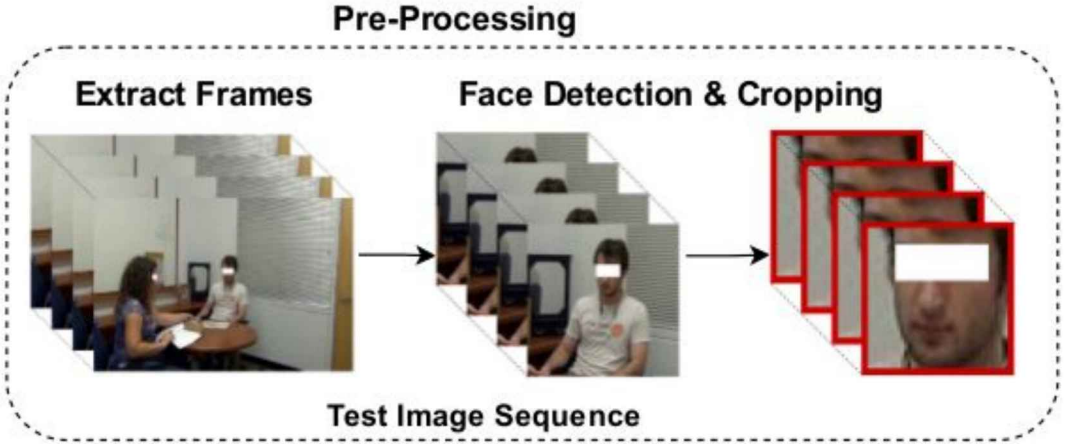
prediction

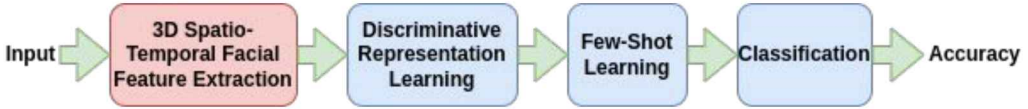


Pre-process

- Split the whole video into 15 separate subvideos based on the 15 activities
- For the chosen scenes (5-7, 11-14)
  - Extract key frames containing the subject of interest
  - Detect and crop square face regions

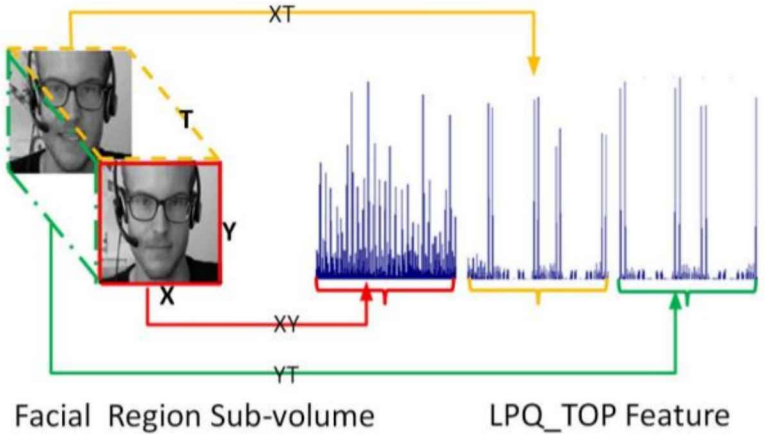






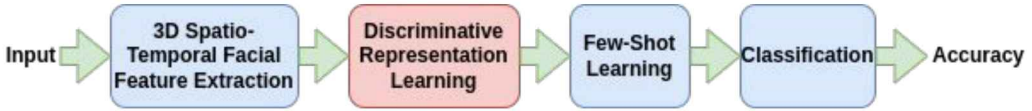
3D Spatio-Temporal Facial Feature Extraction

- LPQ-TOP
  - Local Phase Quantization in Three Orthogonal Planes
- Both spatial and temporal information are extracted
  - spatial
    - facial appearance, static expression, eye movements, etc.
  - temporal
    - expression changes, gaze patterns, and head pose variations



- Basic unit: 2-second long face frames
- 3-dimensional face region subvolume

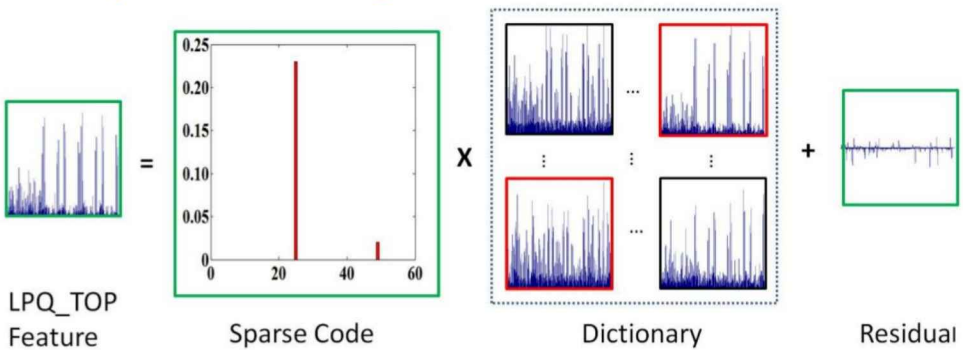
B. Jiang, M. F. Valstar, and M. Pantic, “Action unit detection using sparse appearance descriptors in space-time video volumes,” in Face and Gesture 2011. IEEE, 2011, pp. 314–321.



Discriminative Representation Learning

- Consider a combination of Sparse Coding (K-SVD) and Dimensionality Reduction (MFA)

➔ Sparse Coding



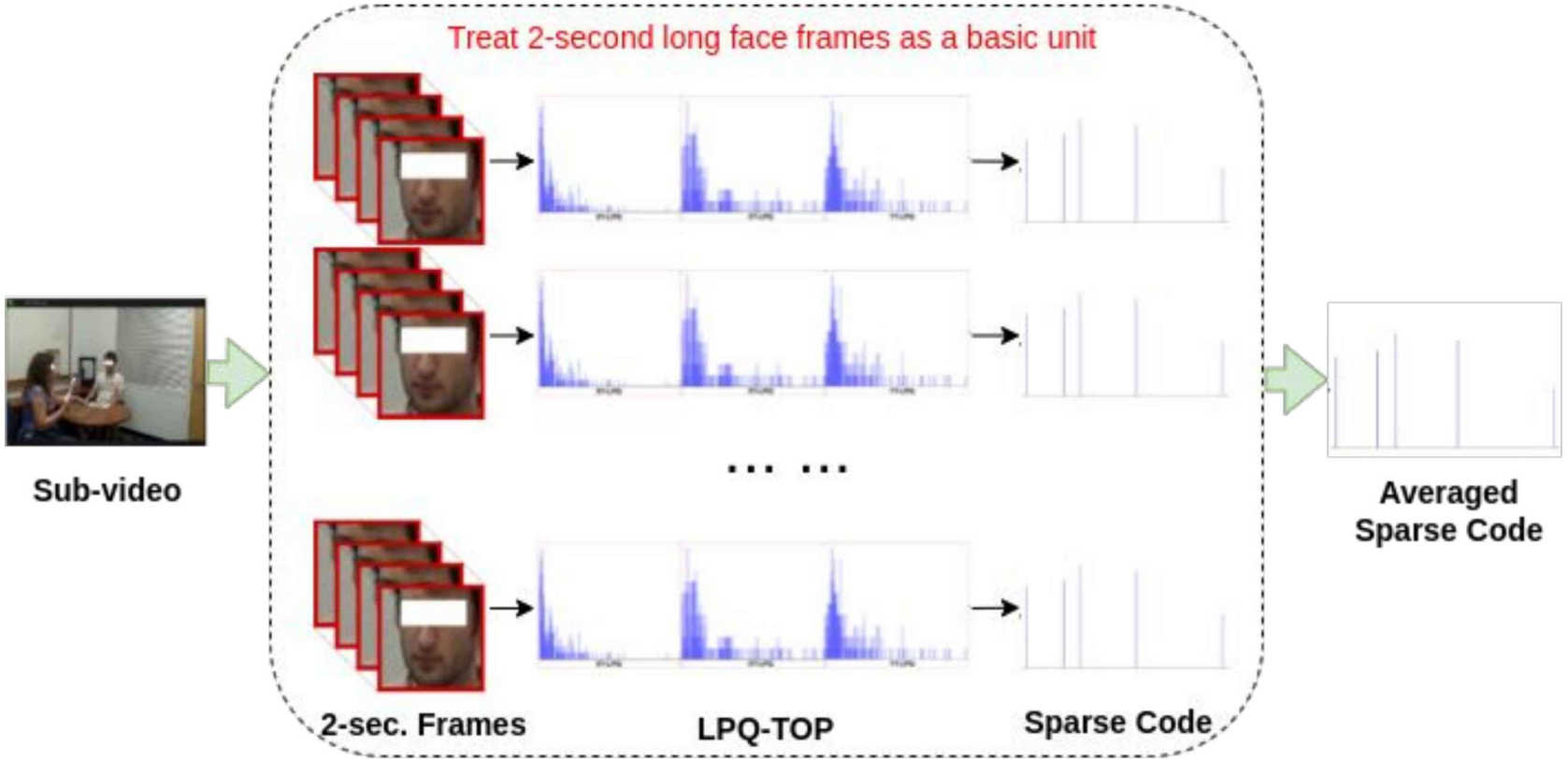
A LPQ-TOP feature is a sparse linear combination of all the dictionary atoms plus a residual or sparse errors

- K-Singular Value Decomposition
  - Learn a dictionary by lowering the **reconstruction error** via sparse coding

$$arg\ min_{D,X} \|Y - DX\|_2^2 \ s.t.\ \forall i,\ \|x_i\|_0 \leq T,$$

  - **T** is a positive integer specifying the sparsity level

M. Aharon, M. Elad, and A. Bruckstein, “K-svd: An algorithm for designing overcomplete dictionaries for sparse representation,” IEEE Transactions on signal processing, 2006.
S. Yan,et al., “Graph embedding and extensions: A general framework for dimensionality reduction,” IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007.
L. Wen, X. Li, G. Guo, and Y. Zhu, “Automated depression diagnosis based on facial dynamic analysis and sparse coding,” IEEE Transactions on Information Forensics and Security, 2015.



After obtain the sparse code of all LPQ-TOP feature in the subvideo, all sparse codes are **averaged** as the descriptor of the subvideo.

➔ MFA: Marginal Fisher Analysis

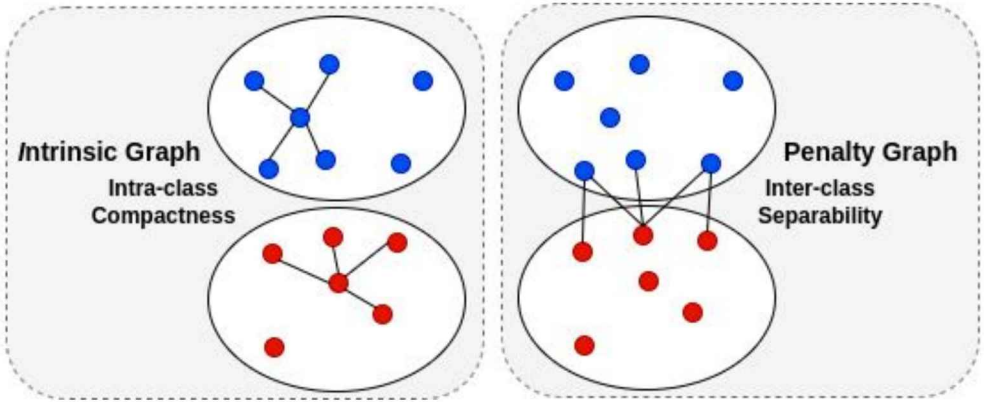
- Map the sparse feature into a new space with better discrimination
- Designs two graphs to characterize:
  - intra-class compactness ⇧
  - inter-class separability ⇧
- Optimize the corresponding equation by obtaining the optimal projection vector v̂:

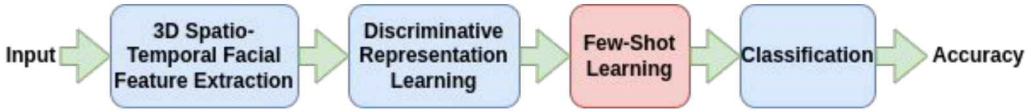
$$\hat{v} = arg\ min_{\mathbf{v}} \frac{v^T X L_{intra} X^T v}{v^T X L_{inter} X^T v}$$

$\mathbf{X} = [x_1, ..., x_n]$ is input data,

$\mathbf{L}_{intra}$ within-class Laplacian matrix

$\mathbf{L}_{inter}$ between-class Laplacian matrix





Few-shot Learning

➔ Distribution Calibration (DC)

- Overfitted if trained on the data with a biased distribution
  - containing only a limited number of samples

- Tukey's ladder of power transformation
  - reduce skewness, more Gaussian-like

$$\tilde{x} = \begin{cases} x^{\lambda} & if\ \lambda \neq 0 \\ log(x) & if\ \lambda = 0 \end{cases}$$

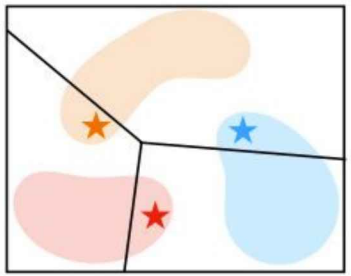
λ: adjust how to correct the distribution.

- Calibrate the mean and the covariance for each class

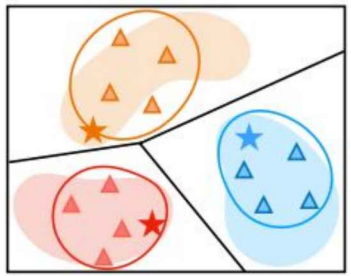
$$\hat{\mu} = \frac{\mu + \tilde{x}}{2}, \hat{\sigma} = \sigma + \alpha$$

α: determines the degree of dispersion of features

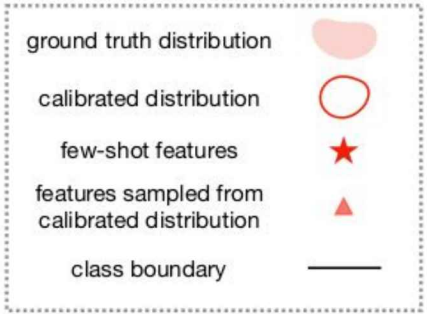
Calibrate the distribution of the few-sample classes by transferring statistics from the classes with sufficient examples



Classifier trained with few-shot features



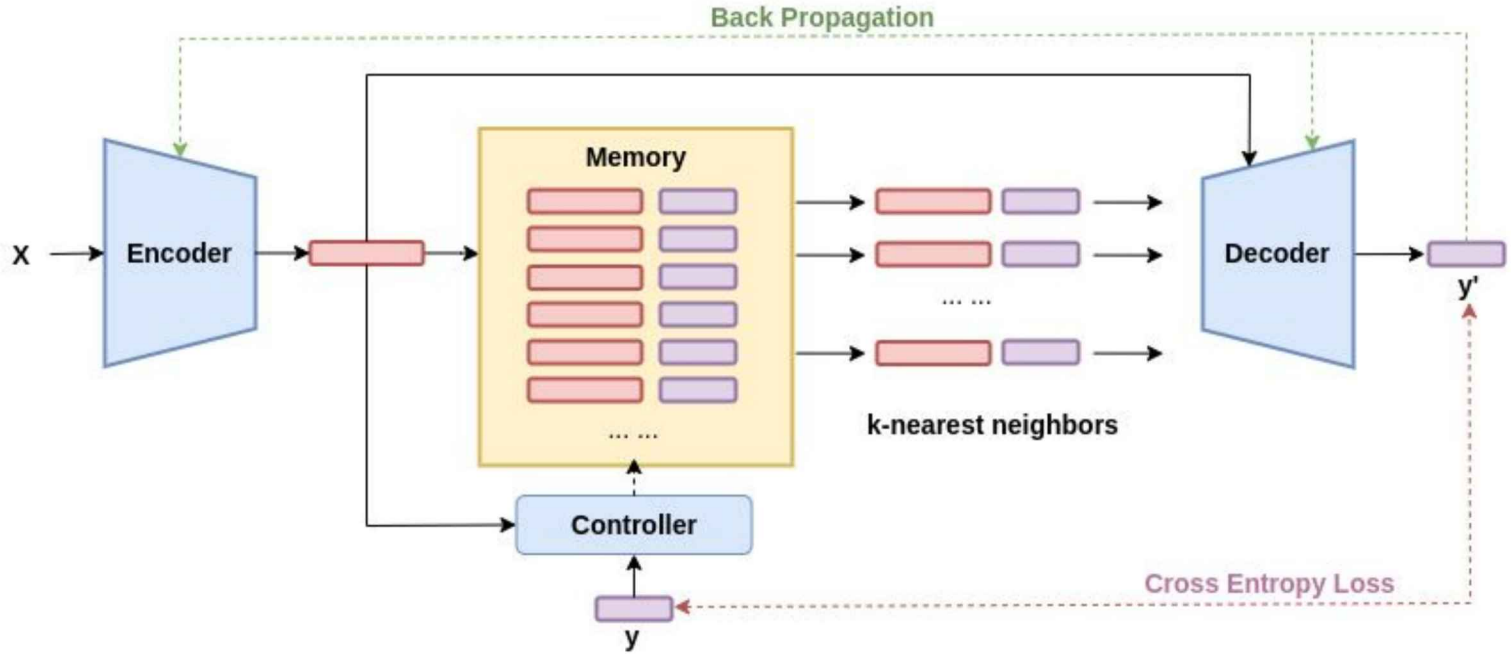
Classifier trained with features sampled from calibrated distribution



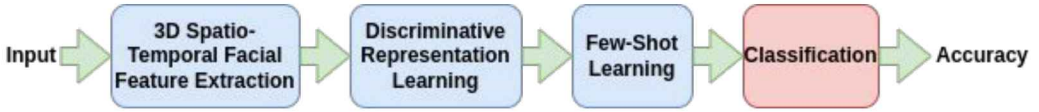
S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," arXiv preprint arXiv:2101.06395, 2021.

➔ Adaptive Posterior Learning (APL)

- Key idea
  - Approximate probability distributions by remembering the most surprising observations it has encountered



T. Ramalho and M. Garnelo, “Adaptive posterior learning: few-shot learning with a surprise-based memory module,” arXiv preprint arXiv:1902.02527, 2019.



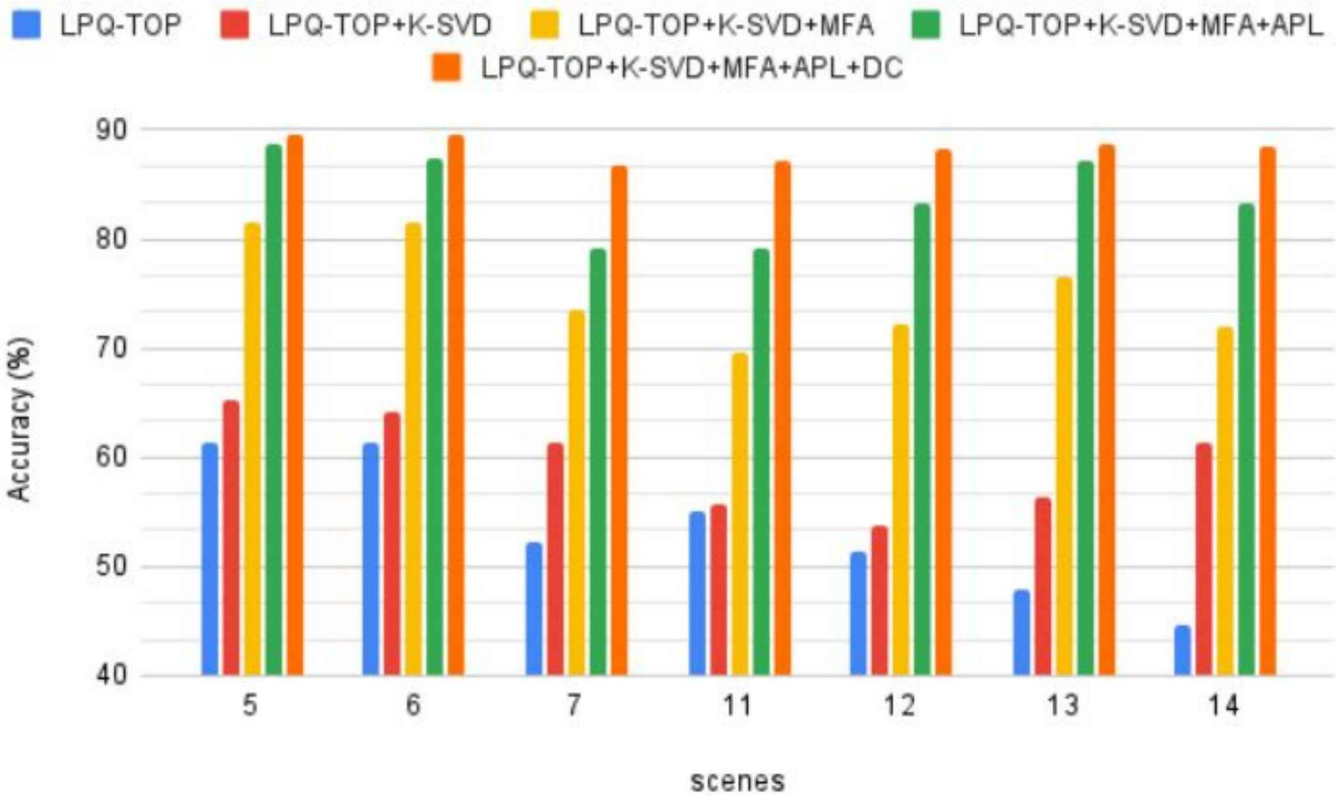
Classification

Scene No.	5	6	7	11	12	13	14
LPQ-TOP	61.40	61.35	52.29	55.11	51.29	47.98	44.72
LPQ-TOP+K-SVD	65.35	64.21	61.26	55.61	53.81	56.35	61.41
LPQ-TOP+K-SVD+MFA	81.45	81.51	73.49	69.58	72.09	76.50	72.00
LPQ-TOP+K-SVD+MFA+APL	88.67	87.35	79.17	79.17	83.33	87.25	83.33
LPQ-TOP+K-SVD+MFA+APL+DC (ours)	**89.64**	**89.55**	**86.83**	**87.12**	**88.33**	**88.67**	**88.49**

- For each scene [5-7, 11-14]
  - Three-class classification
  - 10-fold cross-validation
- Scene-level fusion
  - Top 3, 5, 7

TABLE VI
PERFORMANCE (%) OF OUR METHOD IN FEATURE-LEVEL FUSION (FEATURE CONCATENATION).

Features	Scenes	Accuracy	F1 Score
TOP 3	5,6,13	90.00	86.5
TOP 5	5,6,12,13,14	91.67	90.1
TOP 7	5,6,7,11,12,13,14	**91.72**	90.11


LPQ-TOP
LPQ-TOP+K-SVD
LPQ-TOP+K-SVD+MFA
LPQ-TOP+K-SVD+MFA+APL
LPQ-TOP+K-SVD+MFA+APL+DC
Accuracy (%)
90
80
70
60
50
40
5
6
7
11
12
13
14
scenes