# MM 2021 Tutorial: Few-shot Learning for Multi-Modality Tasks

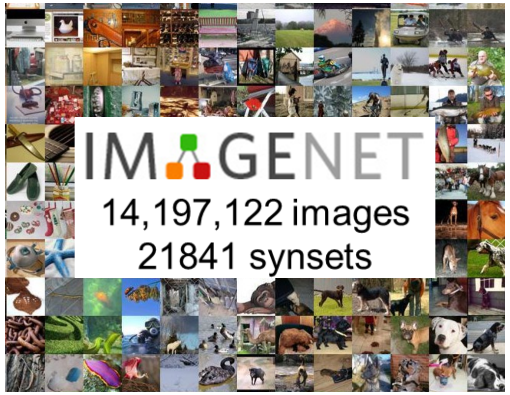
## Multimodal Few-shot and Zero-shot Learning

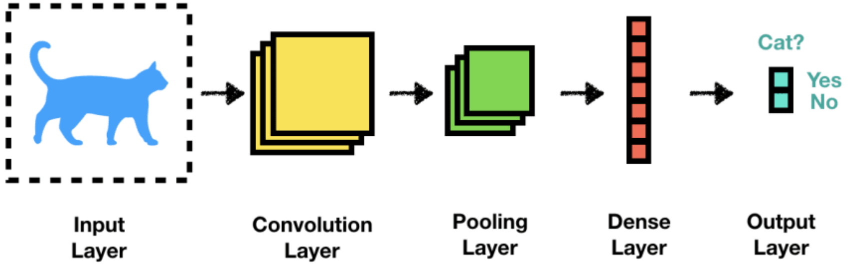

**Xiaoshan Yang**

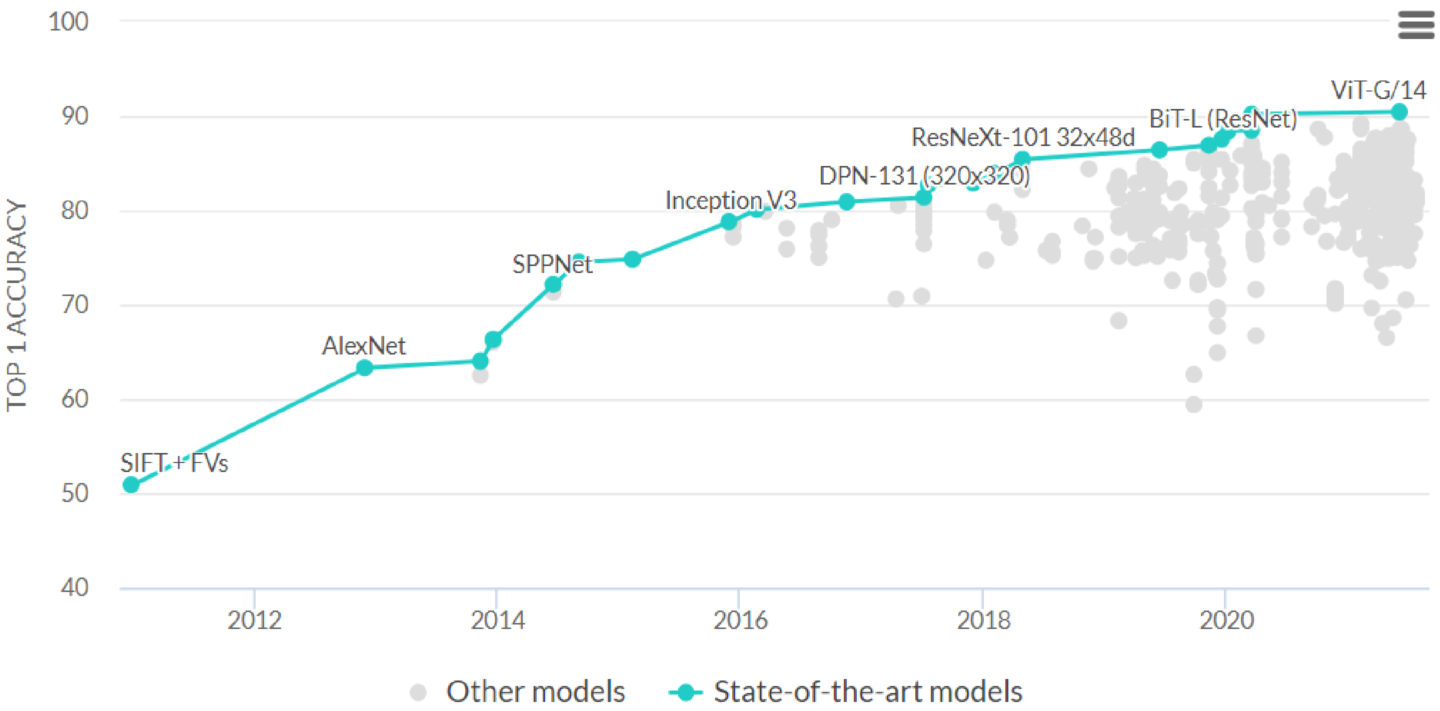
*Associate Prof.* @ Multimedia Computing Group
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences

# Content

# Background



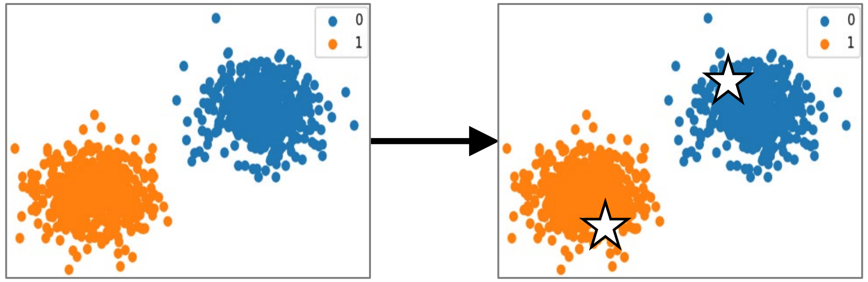




https://paperswithcode.com/sota/image-classification-on-imagenet

**Predict:** $x \to y$

$$(x, y) \sim \Pr[x, y]$$

**Assumption:** Training and test data are from the same distribution.




**Training** **Test**

☆ ☆ **Test samples**

# Background

**Endangered animals**

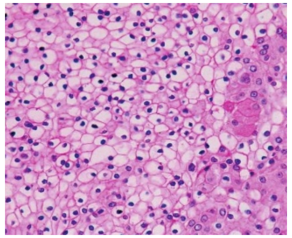
Bombus affinis

Hammerhead shark

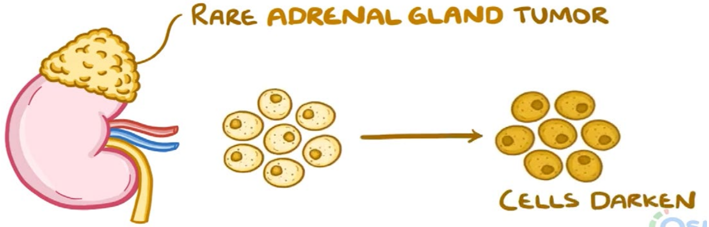
Scimitar oryx

Spider monkey

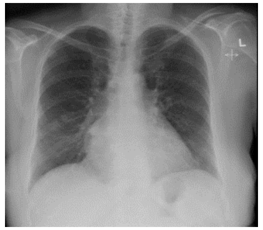
**Orphan tumor lesions**



Chondrosarcoma



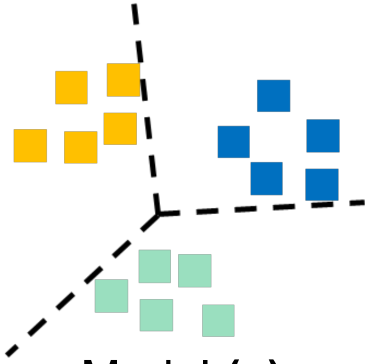

Pheochromocytoma



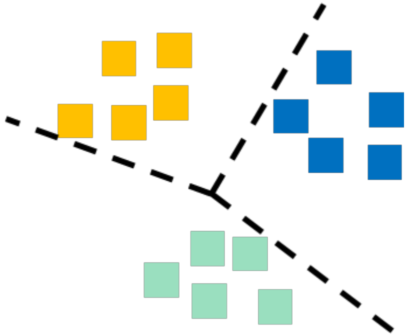
BML

**Rare human actions**

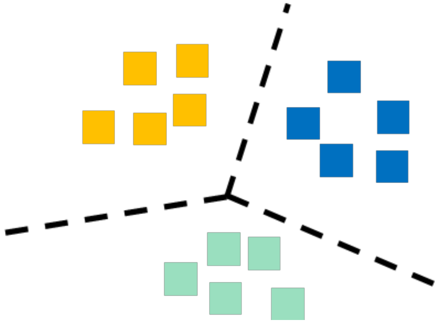
**When there are few samples:**
the deep model suffers serious **overfitting**.



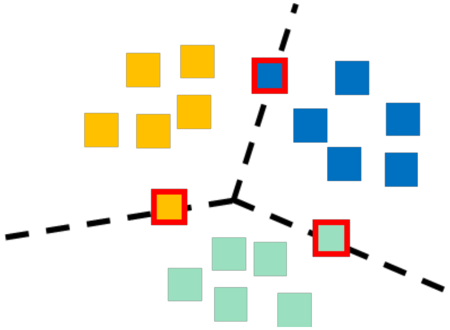
Model **(a)**



Model **(b)**



Model **(c)**



Testing of **(c)**

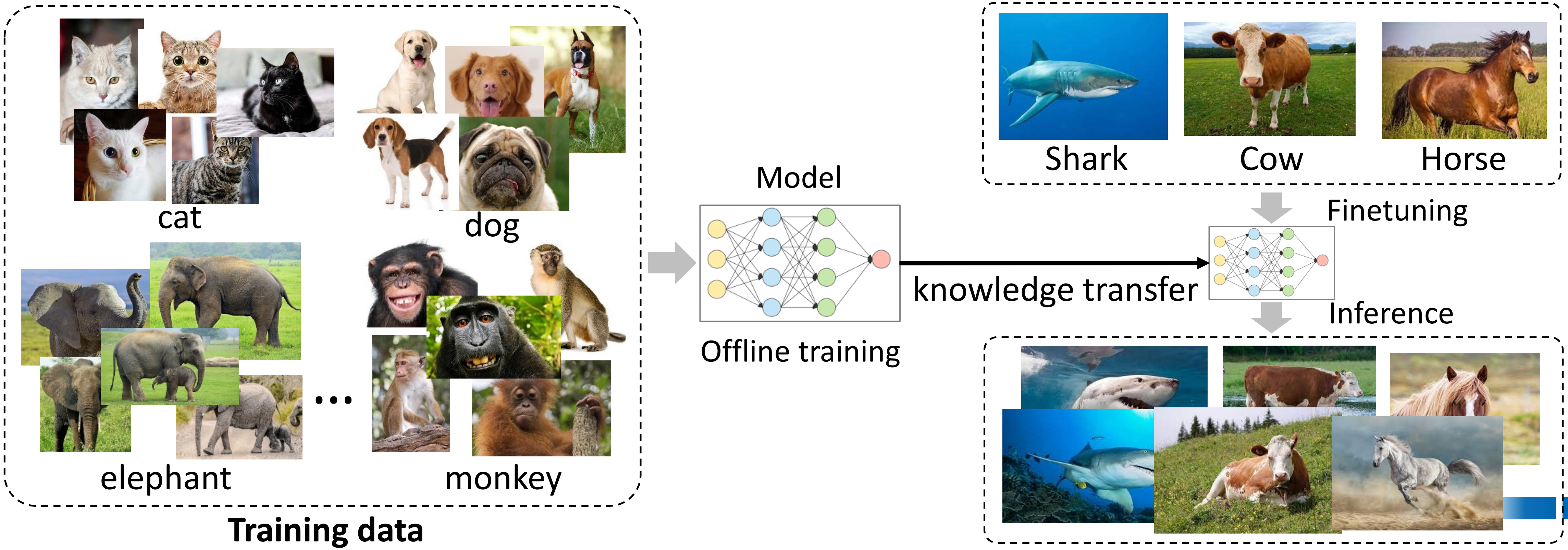
Support samples
Query samples

# Background

➢ Problem definition

- **Few-shot Learning (FSL):** learn a model to recognize novel classes with $K$ support samples per class
- Multimodal FSL: image/text/multimodal data

**Support data** ($N$-way $K$-shot)
(Novel classes with few samples)




**Training data**
(Base classes with abundant samples)

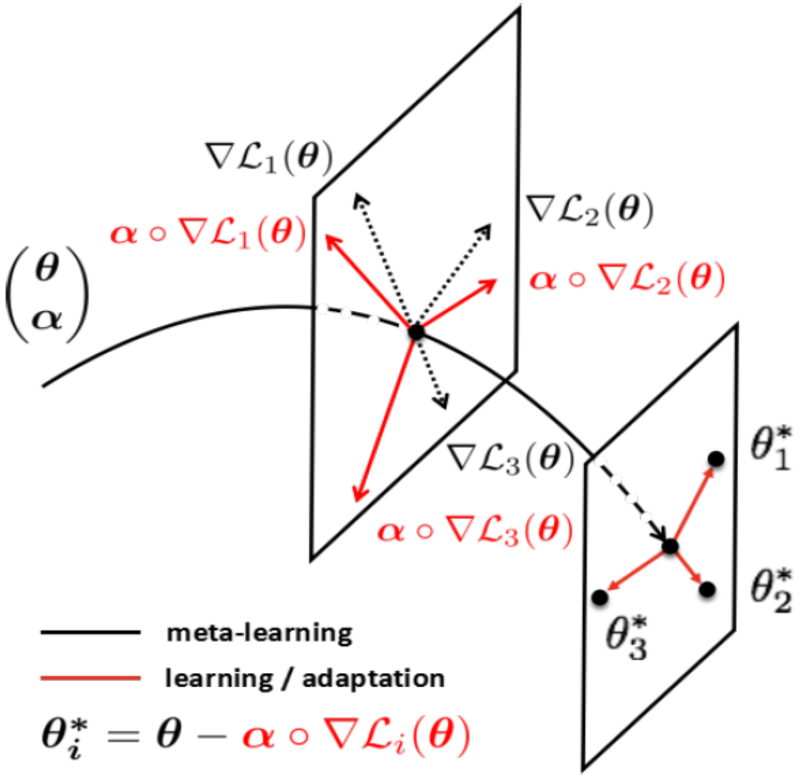
**Query** (test) samples of novel classes

# Background

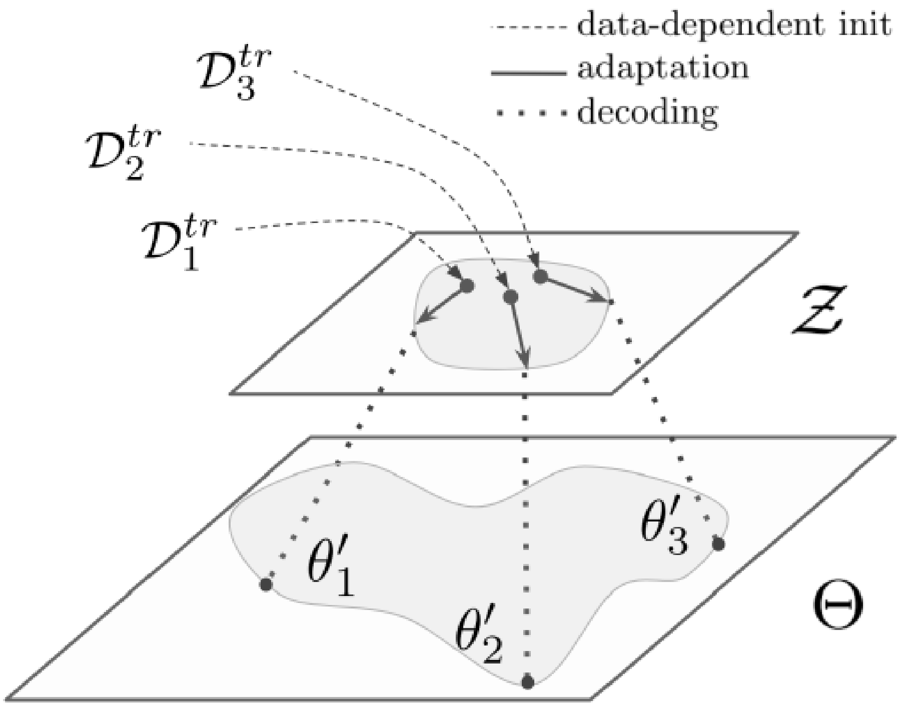
## ➢ Mainstream few-shot learning methods

### ◆ **Gradient-based approaches:**

✓ Optimize the learner to perform well after fine-tuning on the task data done by a single (or few) step(s) of Gradient Descent.




MAML (Model-Agnostic Meta-Learning), ICML 2017



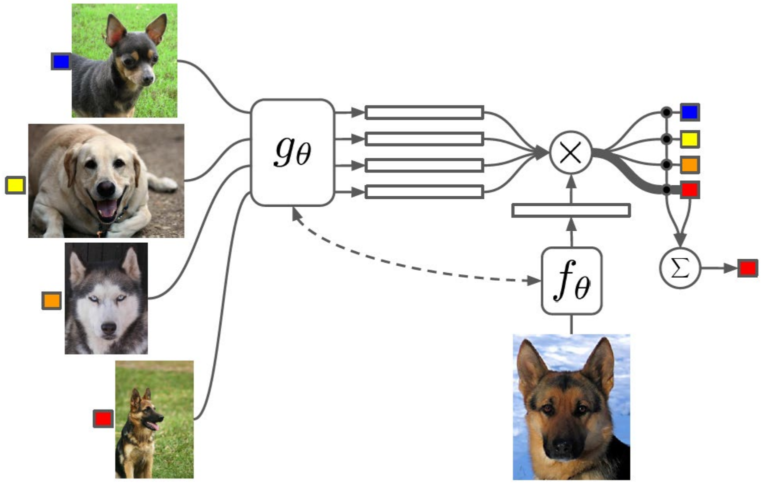

LEO (Latent Embedding Optimization), NeurIPS 2018

# Background

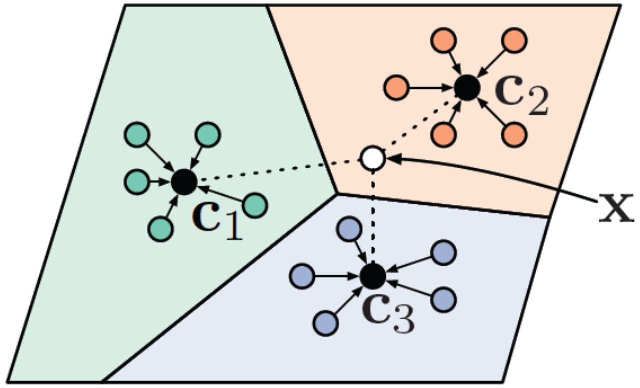
## ➢ Mainstream few-shot learning methods

### ◆ **Metric-based approaches:**

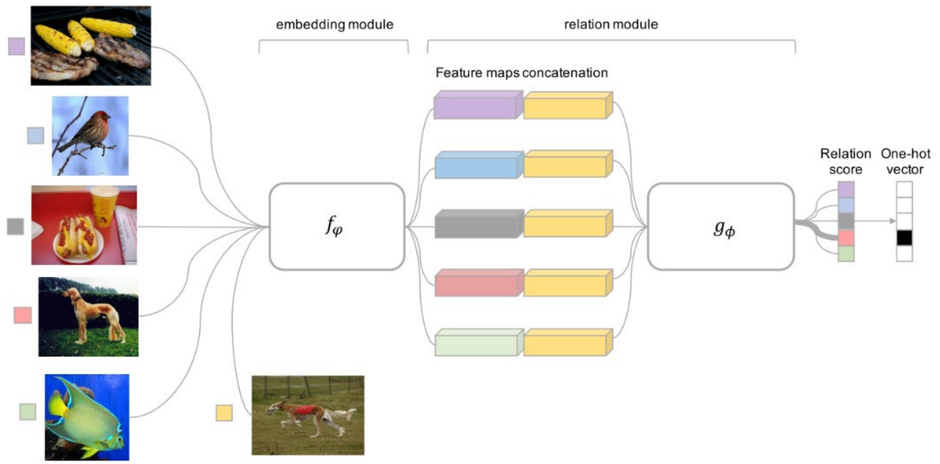
✓ Embed the support and query samples into the same feature space at first, and then compute the similarity of features for prediction.




Matching Networks, NIPS 2016




Prototypical Networks, Nips 2017



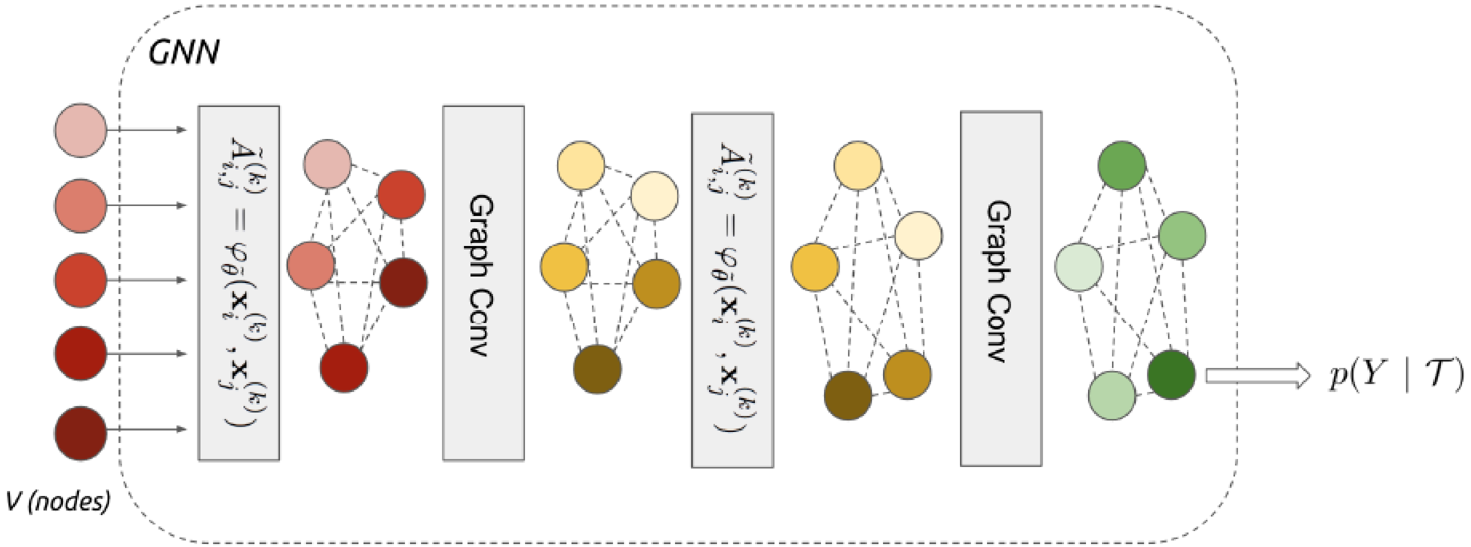

Relation networks, CVPR 2018

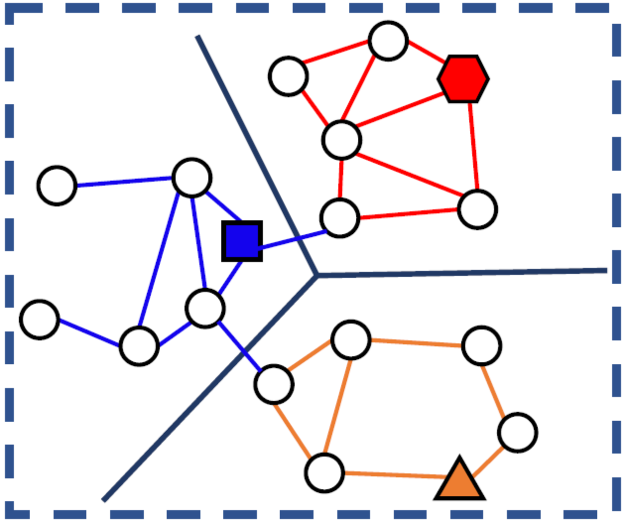
# Background

➢ Mainstream few-shot learning methods

◆ **Metric-based approaches (with GNNs):**

✓ Learn to propagate the class label from the support set to the query set by considering the instance-level relations of samples.



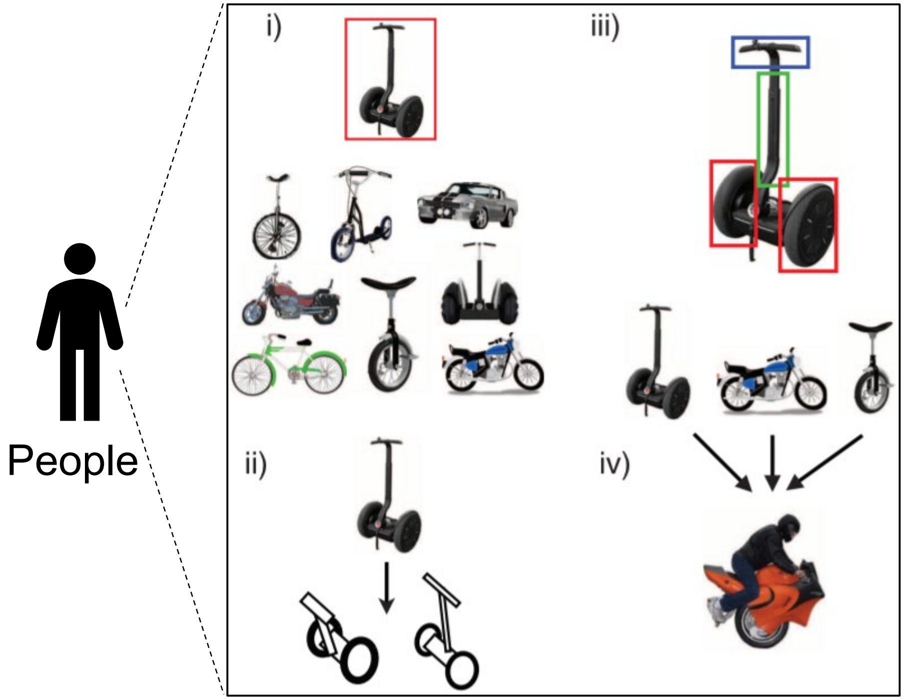
Few-shot Learning with GNNs, ICLR 2018

Transductive Propagation Network, ICLR 2019

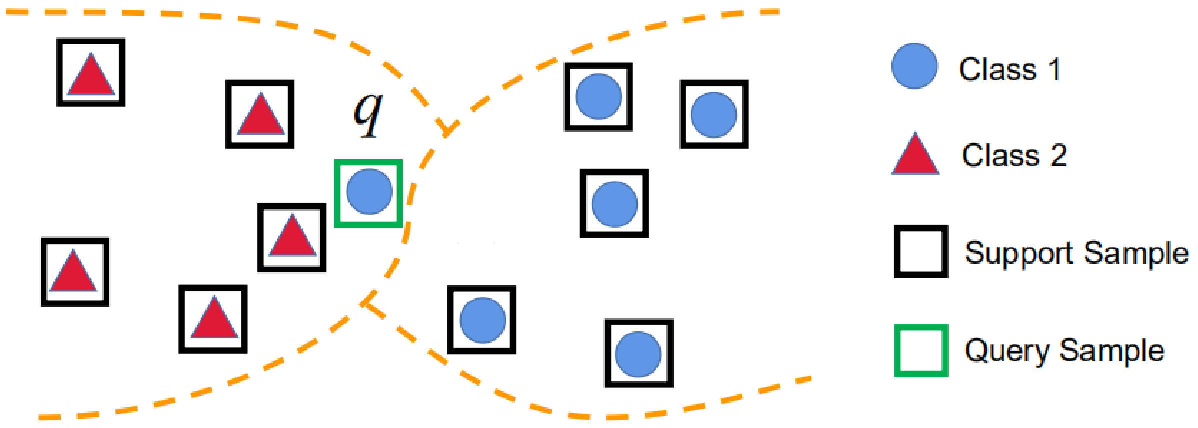
# Explicit Multimodal Knowledge Propagation Network

## ➢ Motivation

- Existing Matching/Relation/Prototype Networks: **sample-sample or sample-category relations**;
- Existing GNNs-based methods: **sample-sample relations**;
- People can learn richer representations of a **new category from just a handful of samples** (**category-sample, sample-sample relations**), using them for creating new exemplars, and even creating new **abstract categories based on existing categories** (**category-category relations**).




**Inspiration: explicitly learn the richer class knowledge to guide the graph-based inference of query samples.**



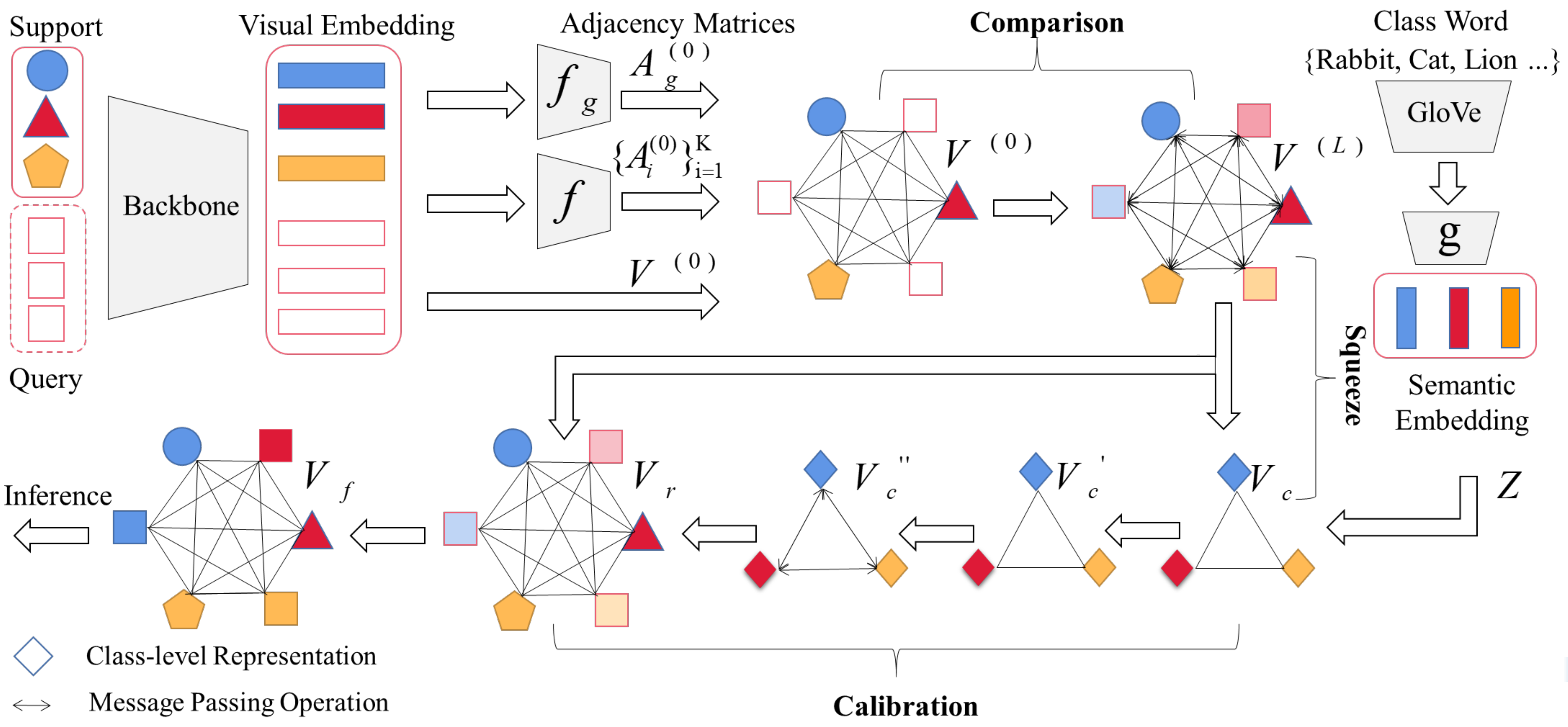

☆ Class-level Knowledge Representation

[Fig]: Human-level concept learning through probabilistic program induction. **Science**, 2015.

# Explicit Multimodal Knowledge Propagation Network

## ➢ Method

- We propose Explicit Class Knowledge Propagation Network (**ECKPN**)
- **Comparison, squeeze and calibration modules** are designed to learn and propagate the class-level knowledge



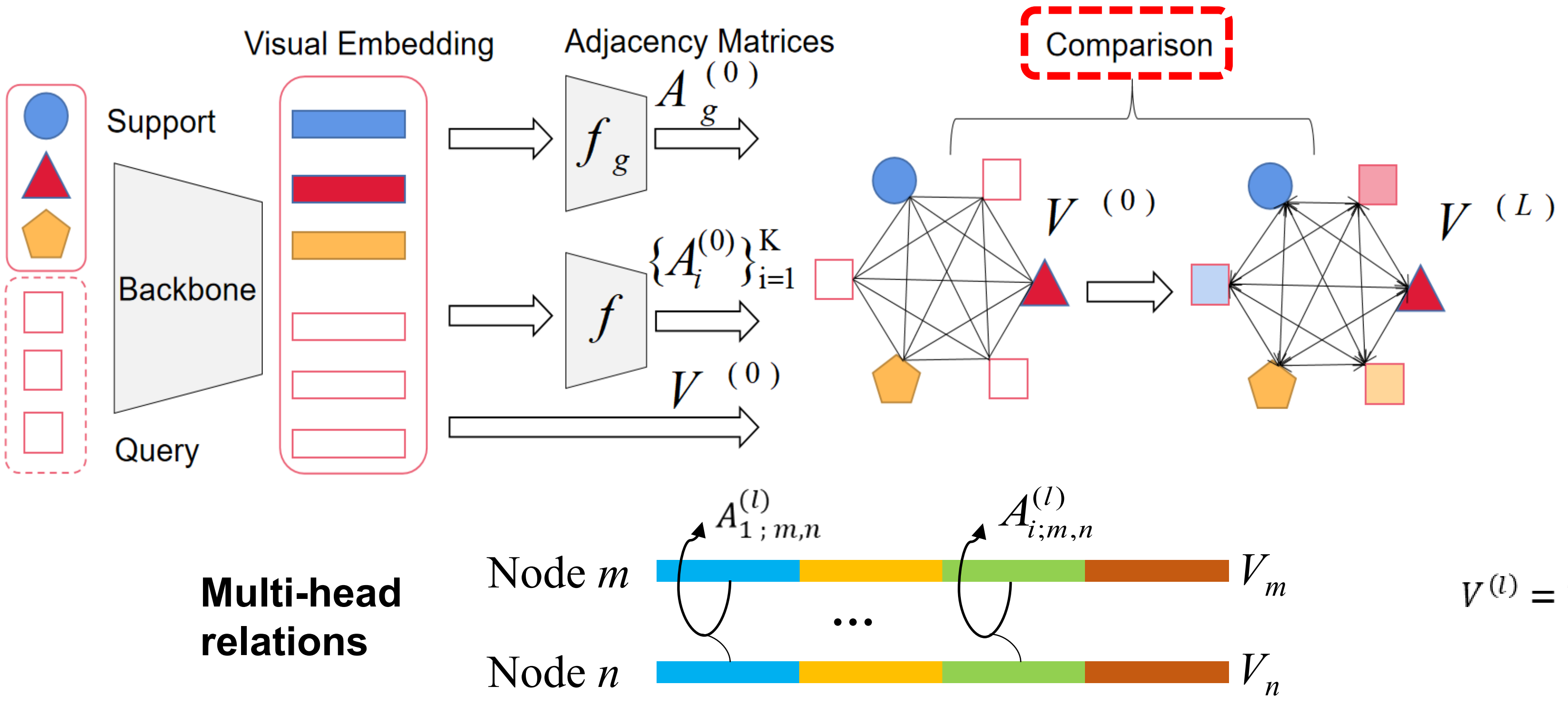

Chaofan Chen, Xiaoshan Yang, Changsheng Xu, Xuhui Huang, Zhe Ma. ECKPN: Explicit Class Knowledge Propagation Network for Transductive Few-shot Learning, ***IEEE CVPR 2021***.

# Explicit Multimodal Knowledge Propagation Network

## ➢ Method

◆ **Comparison Module:** Instance-level Message Passing with Multi-head Relations

- ✓ Build instance-level graph based on support and query samples.
- ✓ Update the sample representations based on the pairwise node relations.
- ✓ Multi-head relations are explored to help model the fine-grained relations of the samples.




Global relations

$$A^{(l)}_{g;m,n} = f_g((V^{(l)}_m - V^{(l)}_n)^2)$$

Multi-head relations

$$A^{(l)}_{i;m,n} = f((V^{(l)}_{i;m} - V^{(l)}_{i;n})^2)$$

Mask of the relations

$$M_{m,n} = \begin{cases} -1 & if\ m,n \in S\ and\ y_m \neq y_n \\ 1 & otherwise \end{cases}$$

Instance-level message passing

$$V^{(l)} = Tr\left(\left[\|_{i=1}^{K}((A_i^{(l-1)} \odot M)V_i^{(l-1)}), (A_g^{(l-1)} \odot M)V^{(l-1)}\right]\right)$$

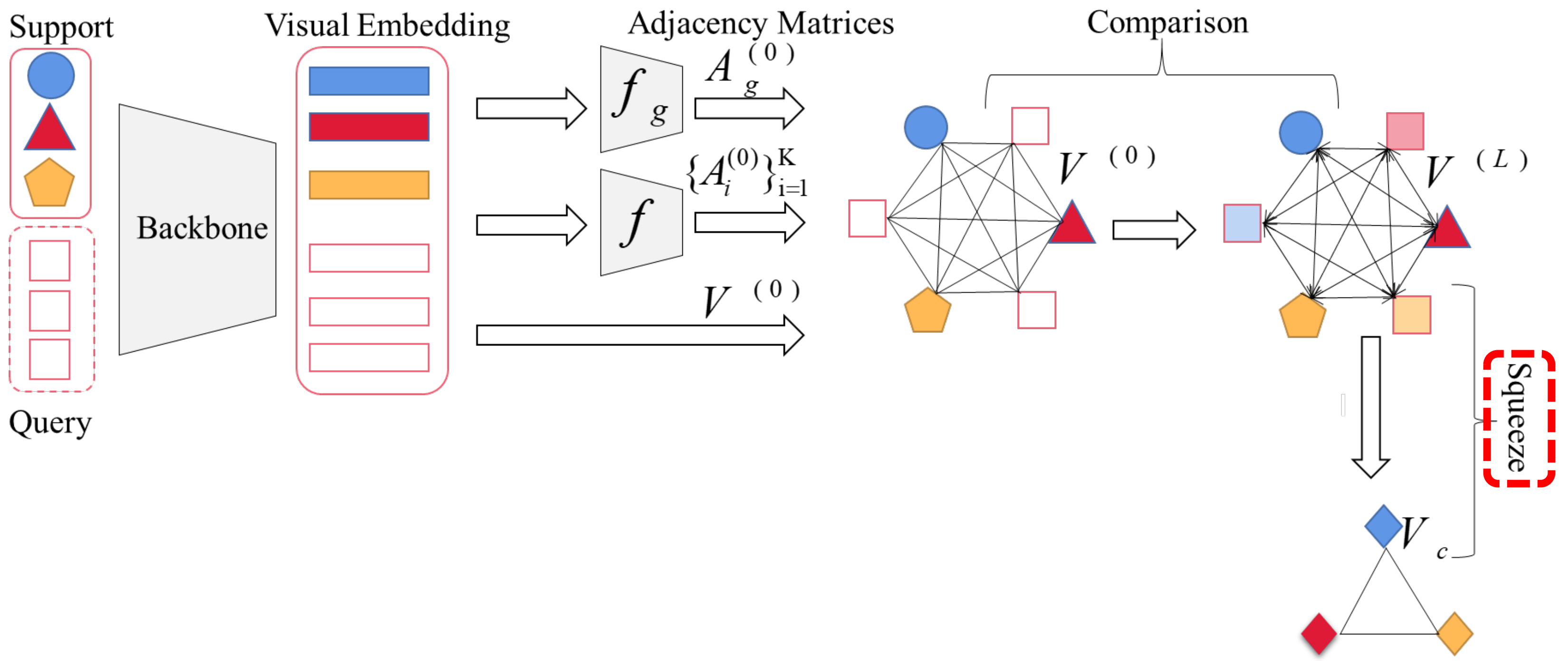
# Explicit Multimodal Knowledge Propagation Network

## ➢ Method

◆ **Squeeze Module:** Class-level Visual Knowledge Learning

✓ Generate the class-level graph based on instance-level graph.

✓ Squeeze samples according to the assignment matrix to obtain the class-level knowledge representations




Assignment matrix

$P = softmax((A_g^{(L)} \odot M)V^{(L)}W)$

Class-level representations

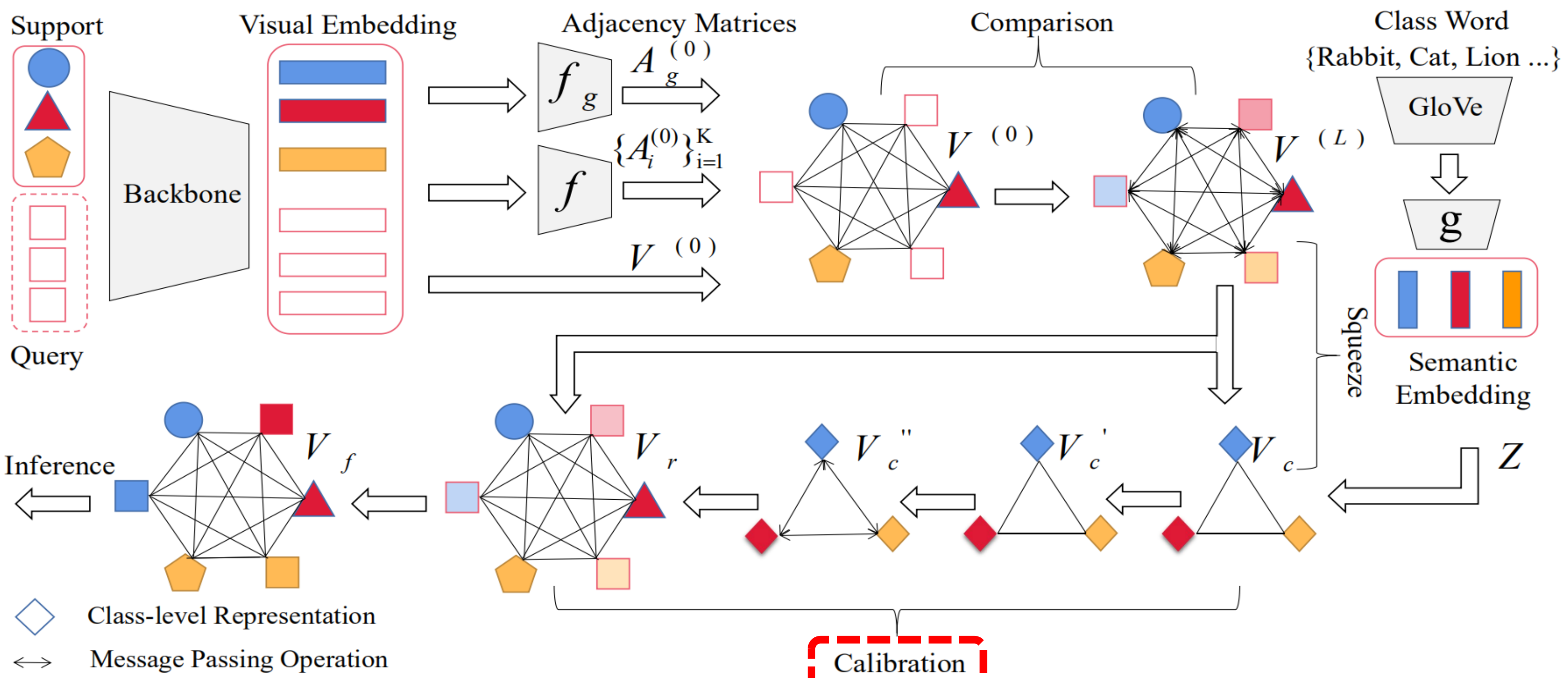
$V_c = P^T V^{(L)}$

# Explicit Multimodal Knowledge Propagation Network

## ➢ Method

◆ **Calibration Module:** Class-level Message Passing with Multi-modal Knowledge

- ✓ Construct multi-modal class knowledge based on word embeddings and visual knowledge.
- ✓ Perform class-level message passing.
- ✓ Combine class-level knowledge representations with instance-level sample representations to guide the inference of the query samples




Multi-modal class representations

$$V_c' = [V_c, Z]$$

Class relations

$$A_c = P^T A_g P$$

Class-level message passing

$$V_c'' = A_c V_c' W'$$

Refined sample representations

$$V_r = P V_c''$$

$$V_f = [V_r, V^{(L)}]$$

Inference of query sample v

$$A_{f;m,n} = f_l((V_{f;m} - V_{f;n})^2)$$

$$\widehat{y_v} = softmax(\sum_{u=1}^{N\times K} A_{f;u,v} \cdot one-hot(y_u))$$

# Explicit Multimodal Knowledge Propagation Network

## ➢ Method

### ◆ Loss Function:

✓ The overall framework is optimized in an end-to-end form with adjacency loss, assignment loss and classification loss.

Adjacency loss
$$\mathrm{L}_0 = -\sum_{A_* \in A_s} \frac{sum(\log(A_*)HG_t)}{sum(HG_t)} + \frac{sum(\log(1-A_*)H(1-G_t))}{sum(H(1-G_t))}.$$

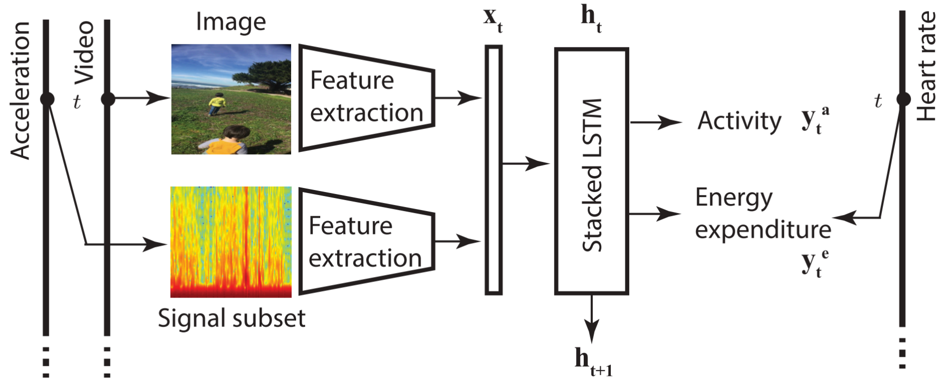
$$A_s = \{A_g^{(1)},\dots,A_g^{(L)}\} \cup \{A_f\} \cup \{A_i^{(1)},\dots,A_i^{(L)}\}_{i=1}^{K} \quad H_{m,n} = \begin{cases} 0 & if\ m \in S \\ 1 & otherwise \end{cases}, \quad G_{t;m,n} = \begin{cases} 1 & if\ y_m = y_n \\ 0 & otherwise \end{cases}$$

Assignment loss
$$\mathrm{L}_1 = \mathrm{L}_{ce}(P, one\text{-}hot([C_s, C_q]))$$

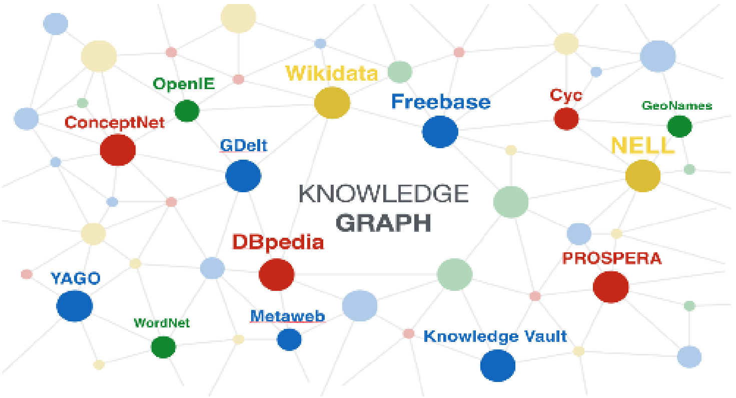
Classification loss
$$\mathcal{L}_2 = \sum_{v \in Q} \mathcal{L}_{ce}(\hat{y}_v, y_v)$$

# Multimodal Few-shot Activity Recognition

- Challenge in egocentric multimodal activity recognition:
  - Modality gap: video ←→sensor signal
  - How to learn an effective activity classifier based on only a few samples per class




- **Motivation:** Knowledge-driven multimodal fusion and activity classification.

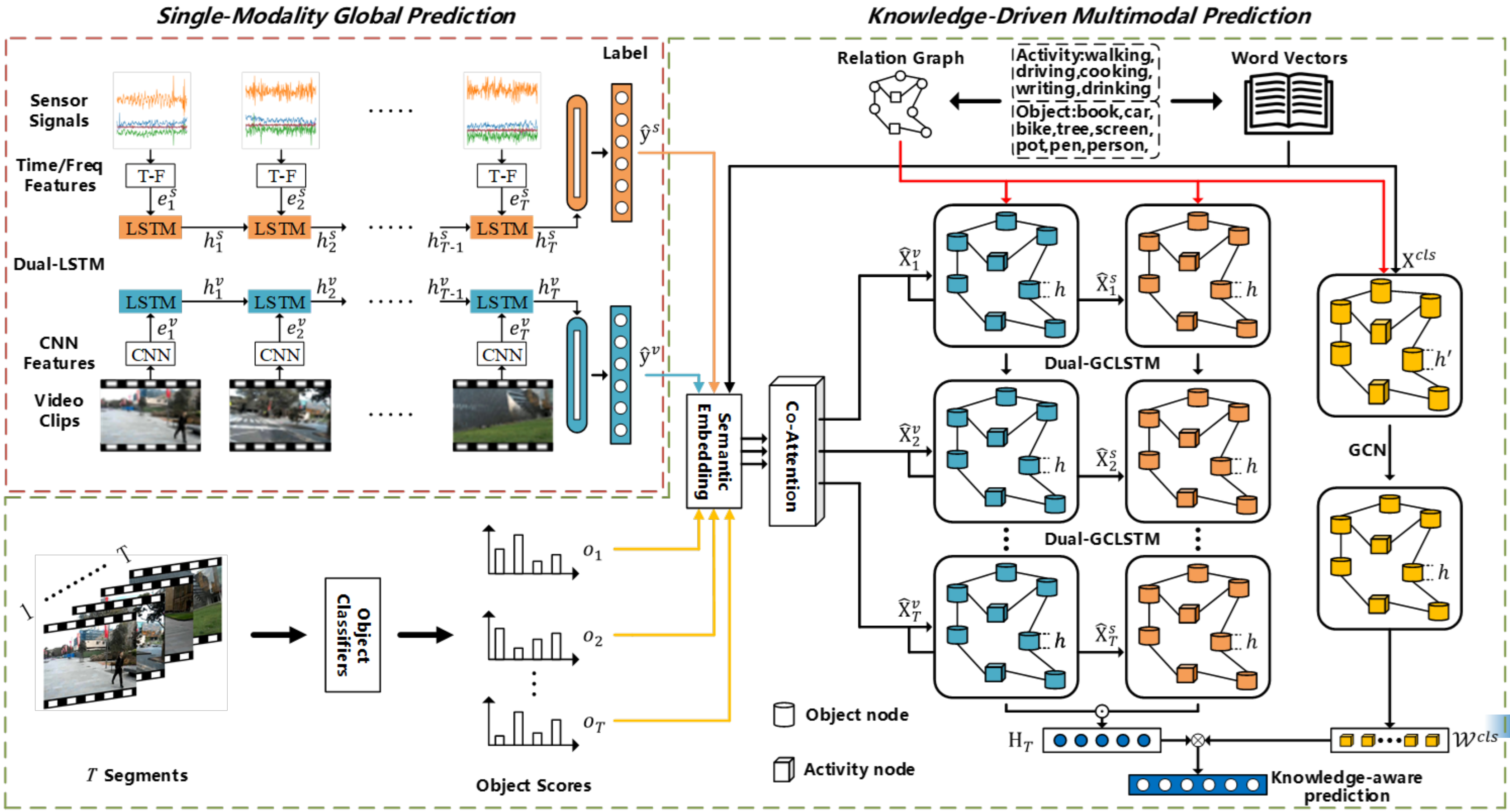





Yi Huang, Xiaoshan Yang, Junyu Gao, Jitao Sang, Changsheng Xu. Knowledge-driven Egocentric Multimodal Activity Recognition, ***ACM TOMM 2021.***

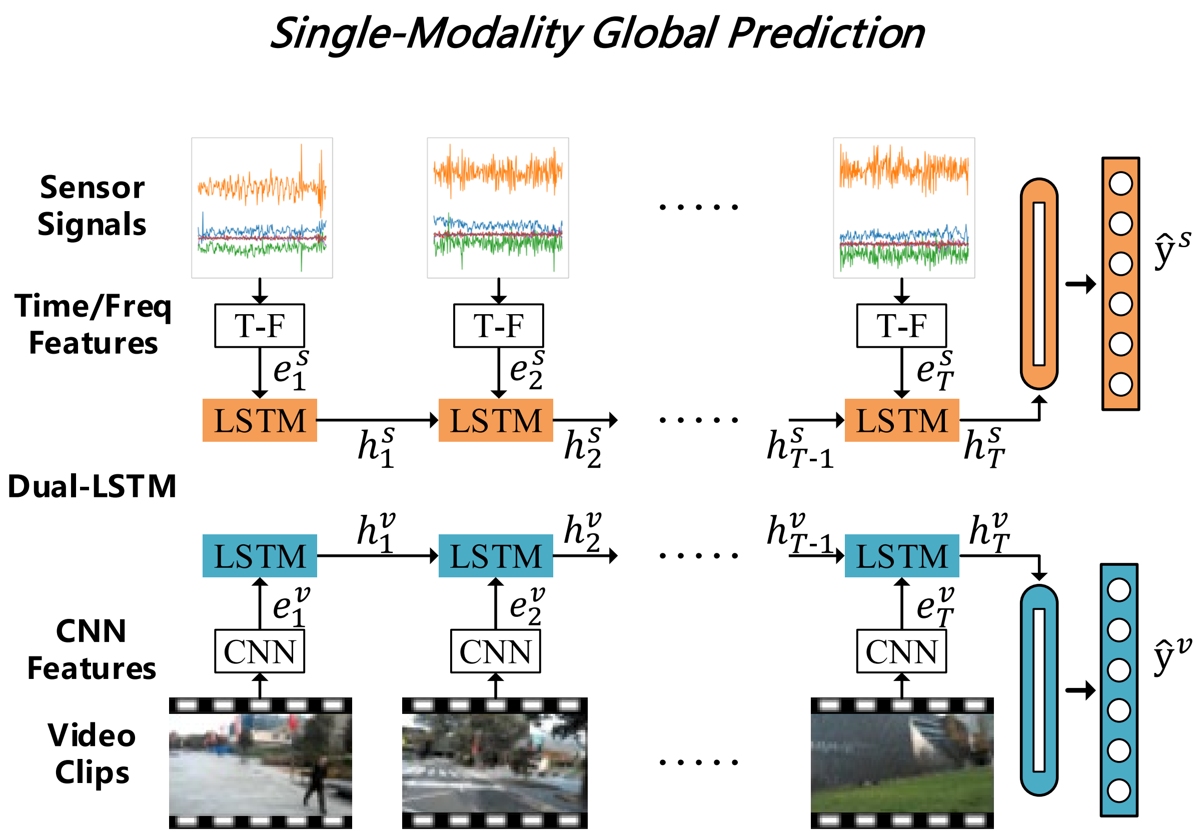
# Multimodal Few-shot Activity Recognition

➢ Method



# Multimodal Few-shot Activity Recognition

- Predicting preliminary activity scores based on single modality




$$\mathbf{h}_t^s = \mathrm{LSTM}_s(\mathbf{e}_t^s, \mathbf{h}_{t-1}^s)$$

$$\hat{\mathbf{y}}^s = \mathrm{Softmax}(\mathbf{W}^s\mathbf{h}_T^s + \mathbf{b}^s)$$

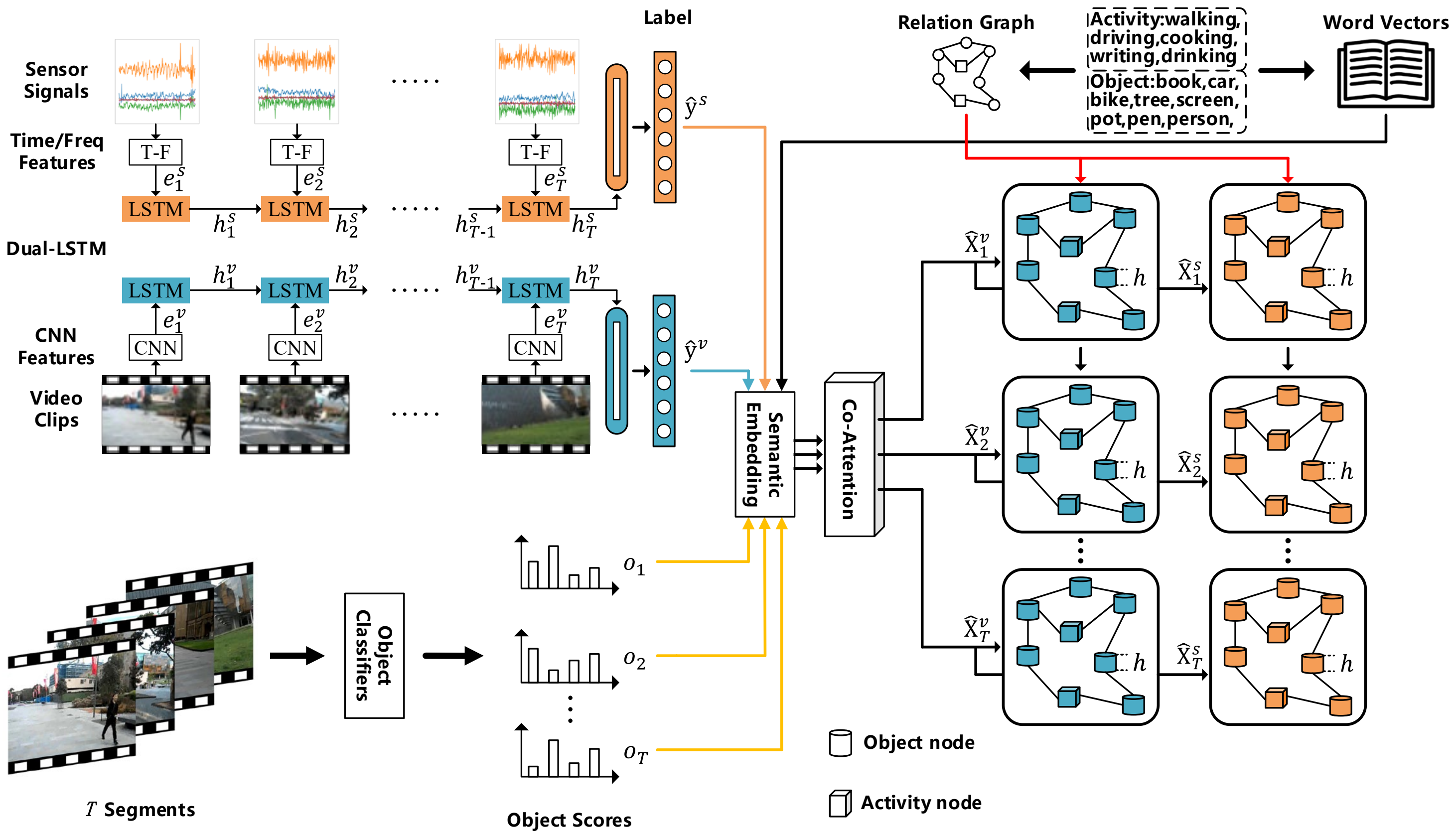
$$\mathbf{h}_t^v = \mathrm{LSTM}_v(\mathbf{e}_t^v, \mathbf{h}_{t-1}^v)$$

$$\hat{\mathbf{y}}^v = \mathrm{Softmax}(\mathbf{W}^v\mathbf{h}_T^v + \mathbf{b}^v)$$

# Multimodal Few-shot Activity Recognition

➢ Dynamic knowledge-aware semantic feature

- ◆ Extracting object as the bridge of video and sensor modalities
- ◆ External knowledge: Relation and semantic embedding of activity and object node

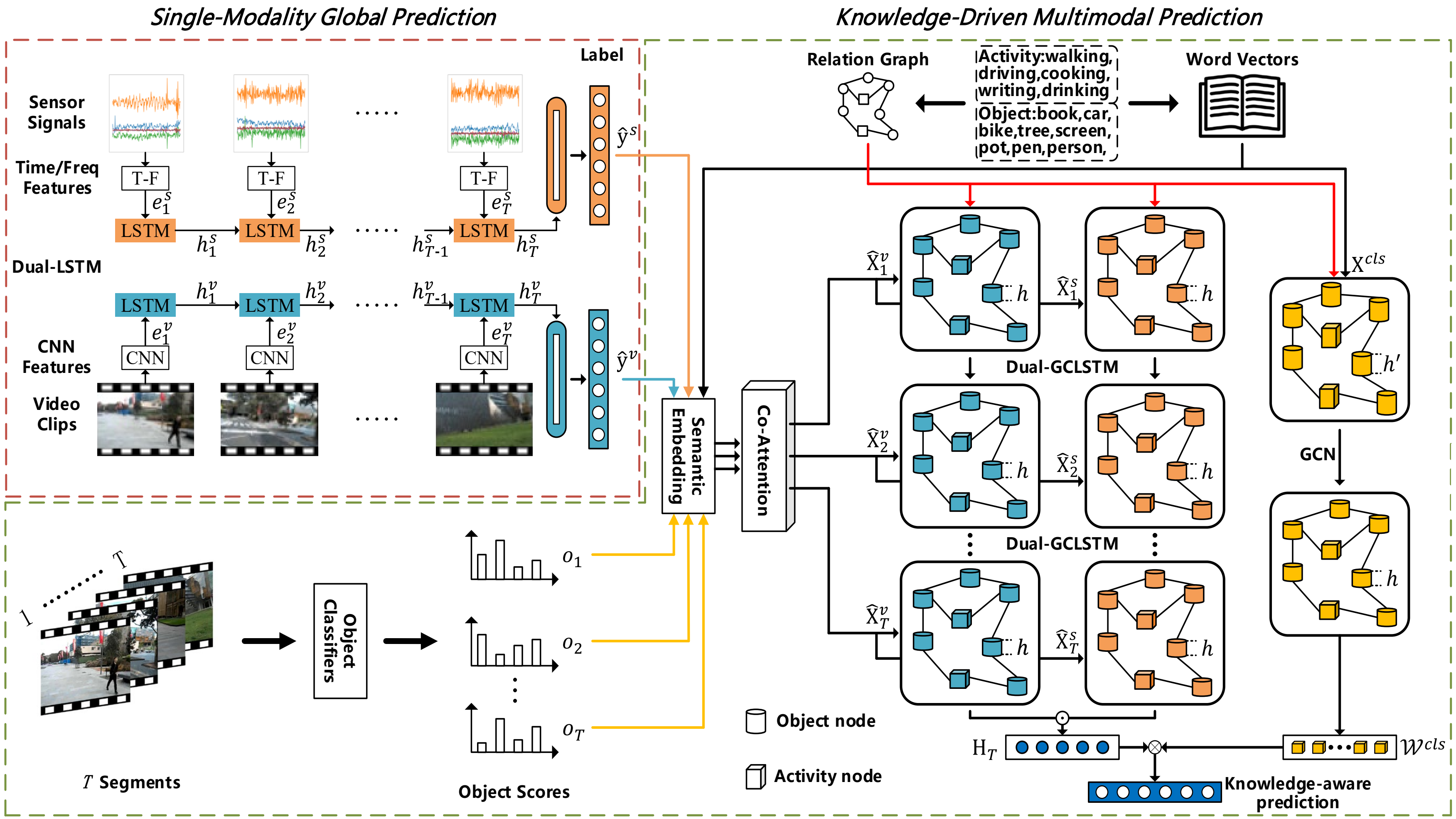



$$\mathbf{X}^{v}_{t,i} = [\hat{\mathbf{y}}^{v}; \mathbf{o}_t]_i \mathbf{X}^{cls}_i$$

$$\mathbf{X}^{s}_{t,i} = [\hat{\mathbf{y}}^{s}; \mathbf{o}_t]_i \mathbf{X}^{cls}_i$$

# Multimodal Few-shot Activity Recognition

➢ Knowledge-aware activity classifier




Knowledge-aware feature

$$\mathbf{H}_t^s = \mathrm{GCLSTM}_s(\hat{\mathbf{X}}_t^s, \hat{\mathbf{H}}_{t-1}^s)$$
$$\mathbf{H}_t^v = \mathrm{GCLSTM}_v(\hat{\mathbf{X}}_t^v, \hat{\mathbf{H}}_{t-1}^v)$$
$$\mathbf{H}_T = \mathbf{H}_T^v \odot \mathbf{H}_T^s$$

Knowledge-aware classifier

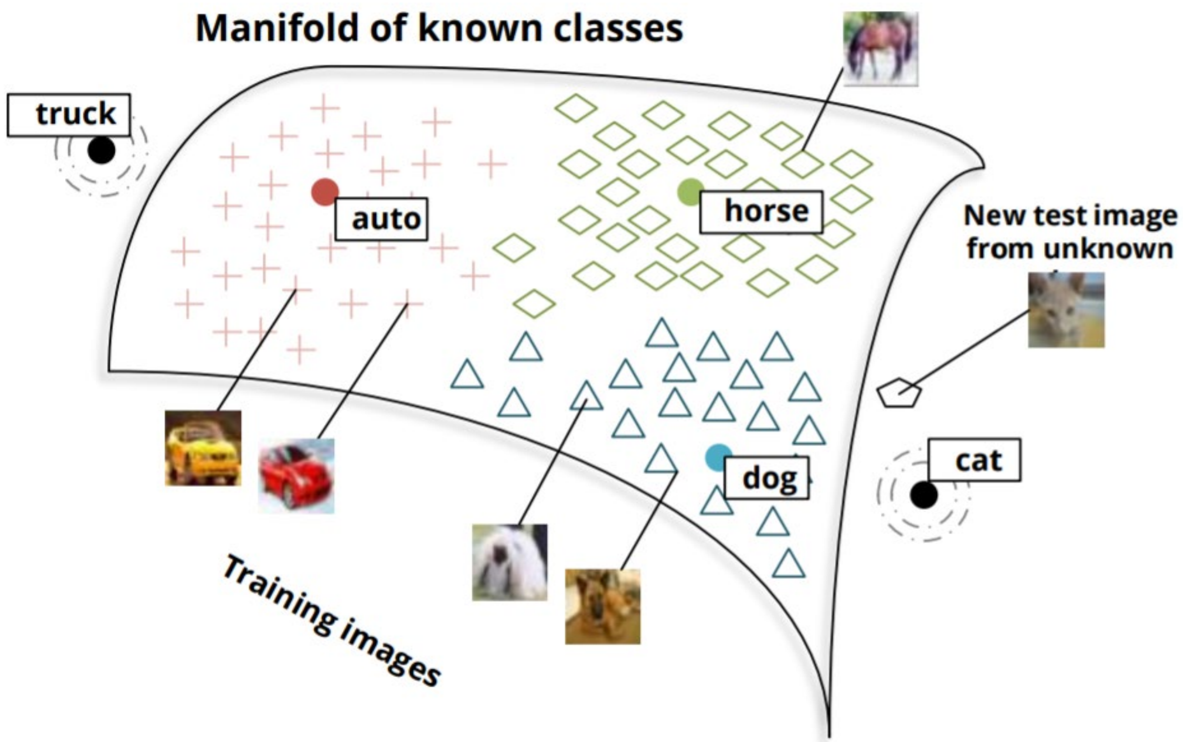
$$\mathcal{W}^{cls} = \mathrm{PReLU}(\mathbf{W}_2 *_{\mathcal{G}} \mathrm{PReLU}(\mathbf{W}_1 *_{\mathcal{G}} \mathbf{X}^{cls}))$$

Knowledge-aware prediction

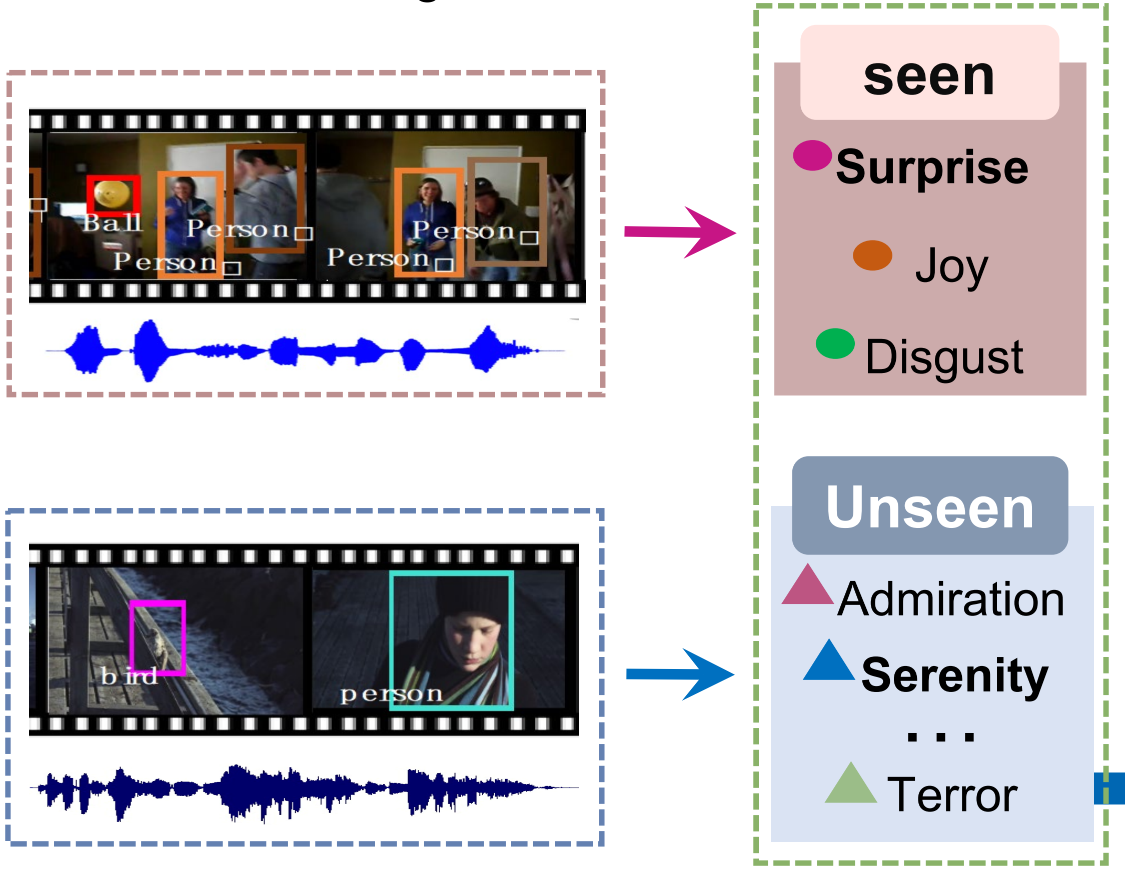
$$\hat{\mathbf{y}} = \mathrm{Softmax}(\mathbf{q})$$
$$q_i = \mathcal{W}_i^{cls} \sum_{o \in \mathcal{N}(i)} \mathbf{H}_{T,o}{}^{\mathrm{T}}$$

# Multimodal Zero-shot Emotion Recognition

- Problem definition
  - **ZSL:** Recognize unseen classes without labeled training instances
  - **One popular solution:** comparing the class description and the sample representation
  - **Multimodal ZSL:** only seen classes have multimodal training data

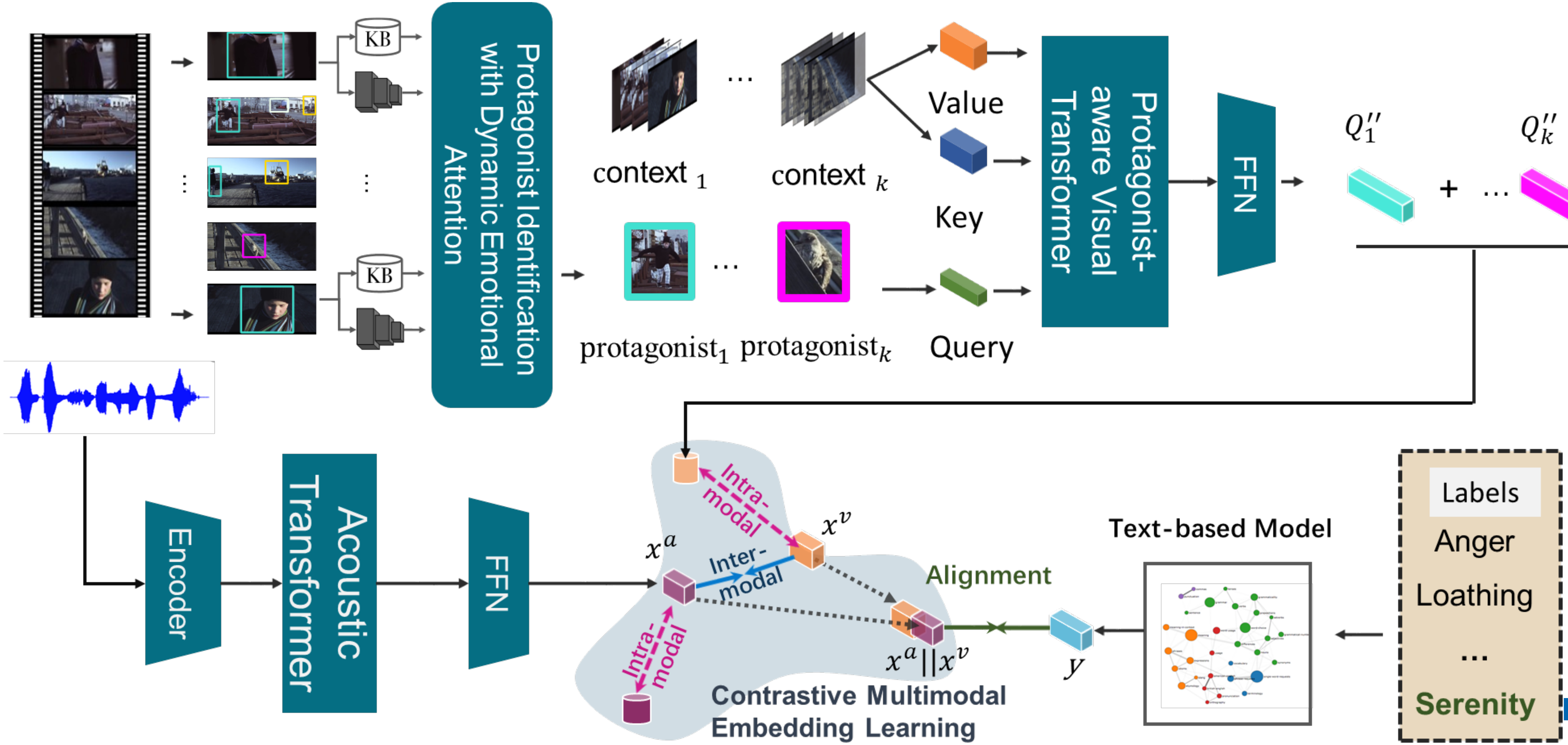



Zero-Shot Learning Through Cross-Modal Transfer, Nips 2013



# Multimodal Zero-shot Emotion Recognition

➢ Method



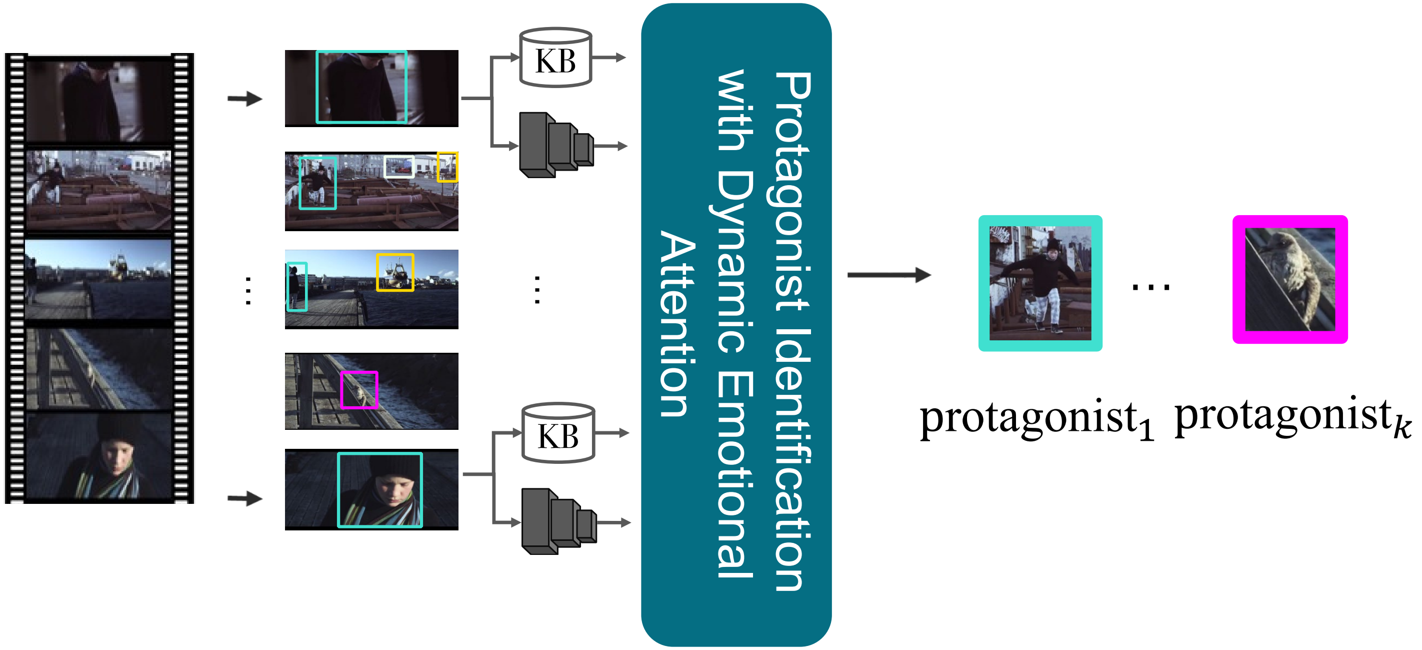



Fan Qi, Xiaoshan Yang, Changsheng Xu. Zero-shot Video Emotion Recognition via Multimodal Protagonist-aware Transformer Network. ***ACM MM 2021***.

# Multimodal Zero-shot Emotion Recognition

## ➢ Method

◆ Protagonist Identification with Dynamic Emotional Attention




$$Q_j = \alpha_j * o_j$$

$$\alpha_j = \sigma(\lambda * rel_j + (1 - \lambda) * aff_j)$$

Relatedness:

$$rel_j = \mathrm{Sum}((conv(o_j, \mathrm{Z}))$$

Affectiveness:

$$aff_j = \left\| \left[ \boldsymbol{V}(\boldsymbol{c}_j) - \frac{1}{2}, \boldsymbol{A}(\boldsymbol{c}_j)/2 \right] \right\|_2$$

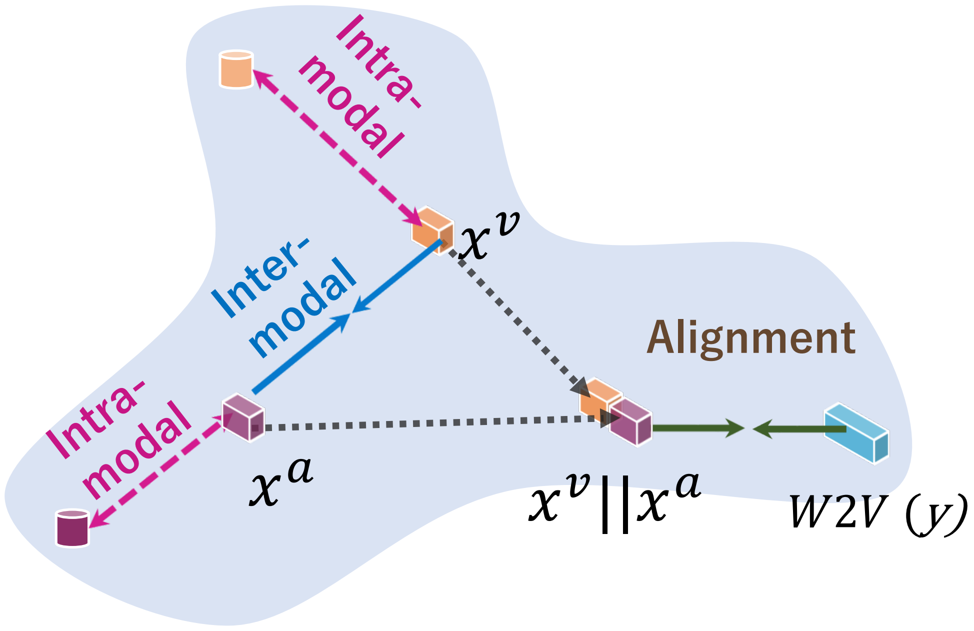
$V$: valence values

$A$: arousal values

# Multimodal Zero-shot Emotion Recognition

## ➢ Method

◆ Contrastive Multimodal Embedding Learning




**Noise Contrastive Estimation Objective(NCE):**

$$l^{NCE}(x^v, x^a) = \frac{\sum e^{s(x^v, x^a)}}{e^{s(x^v, x^a)} + \sum_{(v', a') \sim \mathcal{N}} e^{s(x^{v\prime}, x^{a\prime})}}$$

**Inter-modal NCE:**

$$\mathcal{L}^{inter} = \mathcal{L}(\mathcal{V}, \mathcal{A}) = -\frac{1}{B}\sum_{i=1}^{B} l^{NCE}(x^v, x^a)$$

**Intra-modal NCE:**

$$\mathcal{L}^{intra} = \mathcal{L}(\mathcal{V}, \mathcal{V}^m) + \mathcal{L}(\mathcal{A}, \mathcal{A}^m)$$

**Affective alignment loss:**

$$\mathcal{L}^{al}(\mathcal{V}, \mathcal{A}, \mathcal{Y}) = -\frac{1}{B}\sum_{i=1}^{B} ||x_i^v||x_i^a - W2V(y_i)||_2^2$$

# Conclusion

- Modeling multi-head relations among different samples in GNN-based FSL is simple but effective, it deserves more research work.

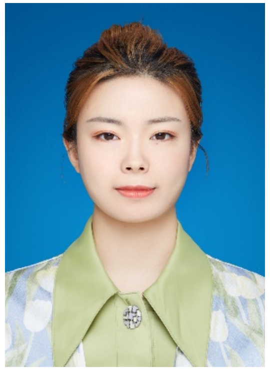
- Using semantic relation knowledge of classes is not as important as in zero-shot learning, we need to explore more complementary knowledge for visual features.

- For multimodal few-shot/zero-shot learning task， we firstly fuse the multimodal features and then apply existing few-shot/zero-shot learning approaches. Unresolved questions：
    - What is the difference between early fusion and late fusion in few-shot/zero-shot case ?
    - Whether it is necessary to use all modalities for each sample in the few-shot/zero-shot learning task ?

# Acknowledgement

- Chaofan Chen, Xiaoshan Yang, Changsheng Xu, Xuhui Huang, Zhe Ma. ECKPN: Explicit Class Knowledge Propagation Network for Transductive Few-shot Learning, ***IEEE CVPR 2021***.
- Yi Huang, Xiaoshan Yang, Junyu Gao, Jitao Sang, Changsheng Xu. Knowledge-driven Egocentric Multimodal Activity Recognition, ***ACM TOMM 2021***.
- Jinxing Pan, Xiaoshan Yang, Yi Huang, Changsheng Xu. Few-shot Egocentric Multimodal Activity Recognition, ***ACM MM Asia 2021***.
- Fan Qi, Xiaoshan Yang, Changsheng Xu. Zero-shot Video Emotion Recognition via Multimodal Protagonist-aware Transformer Network. ***ACM MM 2021***.

*Prof.* Changsheng Xu
*NLPR, CASIA, UCAS*



*Dr.* Fan Qi
*Tianjin University of Technology*

Chaofan Chen
*PHD Candidate, USTC*

Yi Huang
*PHD Candidate, NLPR, CASIA*

Jinxin Pan
*Master, Hefei University of Technology*

# Thanks for Your Attention

xiaoshan.yang@nlpr.ia.ac.cn
https://yangxs.ac.cn