腾讯优图

CAPro: Webly Supervised Learning with Cross-Modality Aligned Prototypes

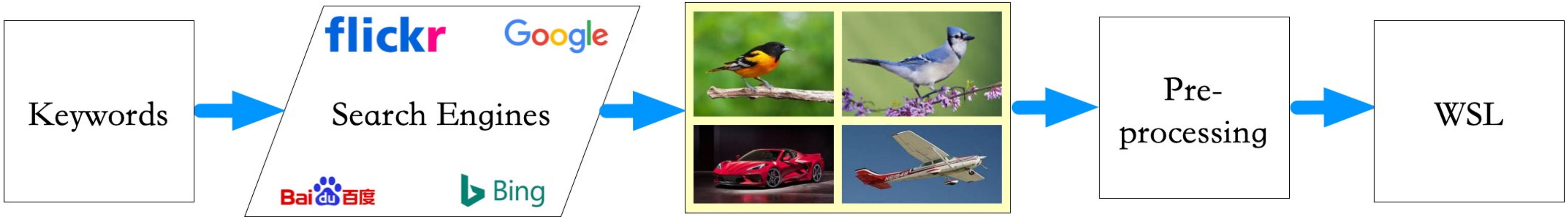
Yulei Qin, Xingyu Chen, Yunhang Shen, Chaoyou Fu,
Yun Gu, Ke Li, Xing Sun, Rongrong Ji

NeurIPS 2023

Tencent YouTu Lab

Webly-Supervised Learning (WSL)

- How to learn noise-robust representations of visual concepts from web data?
  - Various types of noise, especially **the semantic noise**, are under-explored.
  - **Self-bootstrapping** on each sample is prone to overfitting.



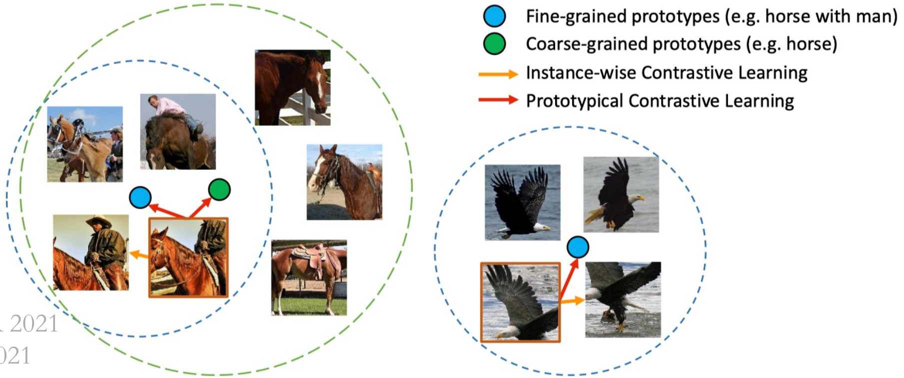
Keywords: Drumsticks (Instrument)

Keywords: Nail (Metal Fastener)

Preliminary: MoPro

腾讯优图

- Prototypes
  - representative embeddings for a group of **semantically similar** instances
- Contrastive Learning
  - **self-supervised** learning method
  - samples from the **same** instance **closer**
  - samples from **different** instances **farther**
- Momentum
  - smooth and consistent optimization policy for **prototypes** and **networks**

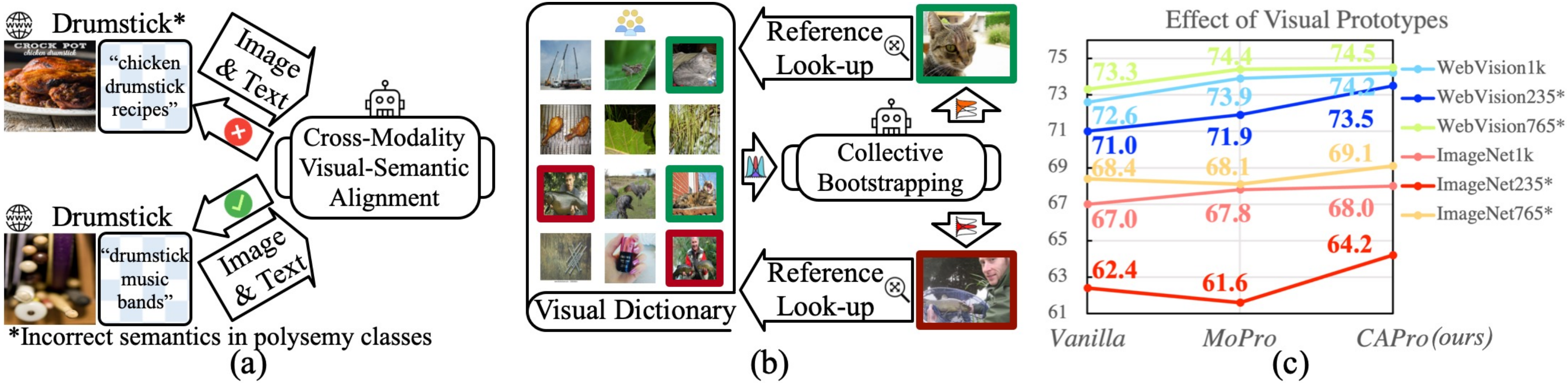


Prototypical Contrastive Learning of Unsupervised Representations, ICLR 2021
Learning from Noisy Data with Robust Representation Learning, ICCV 2021
MoPro: Webly-Supervised Learning with momentum Prototypes. ICLR 2021

Cross-modality Aligned Prototypes (CAPro)

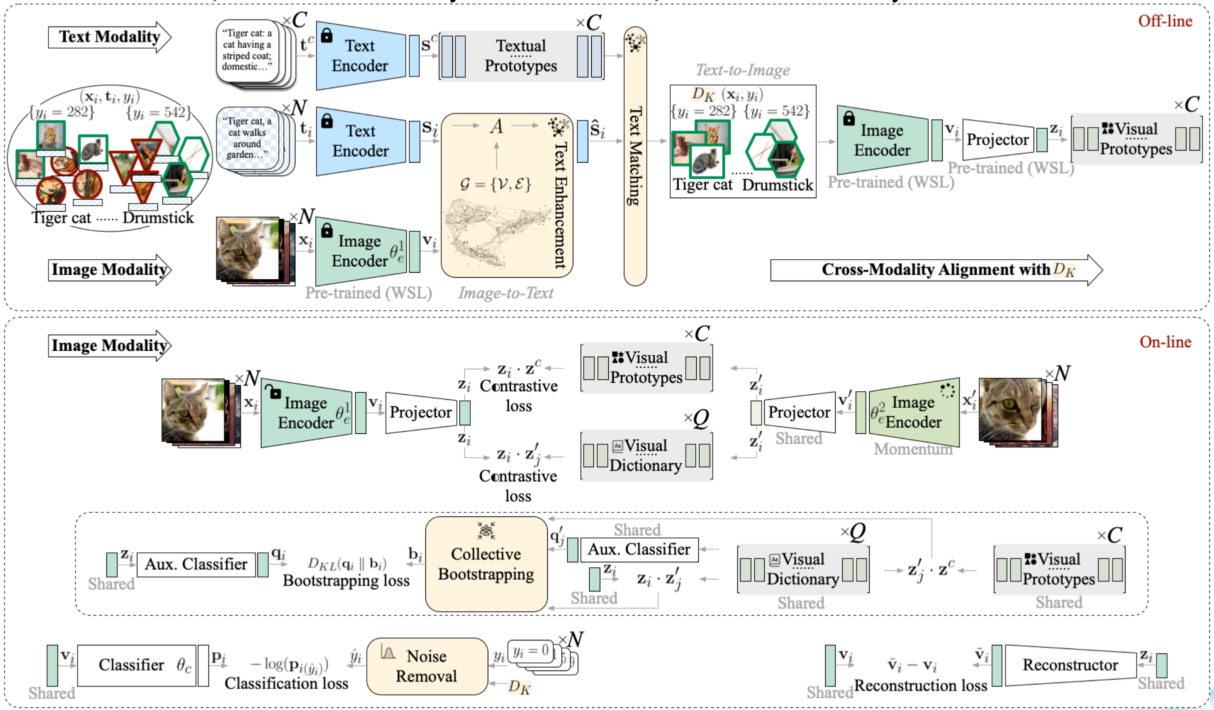
腾讯优图

- What can we do with multi-modal web data (images and texts) for WSL?
- Our contributions:
  - Cross-modality alignment to formulate semantically-correct **textual and visual prototypes**
  - Collective bootstrapping to provide wiser, smoother labels with **collective knowledge**



Overall Model Architecture

- Model components
  - Siamese image encoders, a text encoder, a classifier, a projector, a reconstructor, an auxiliary classifier, a dictionary

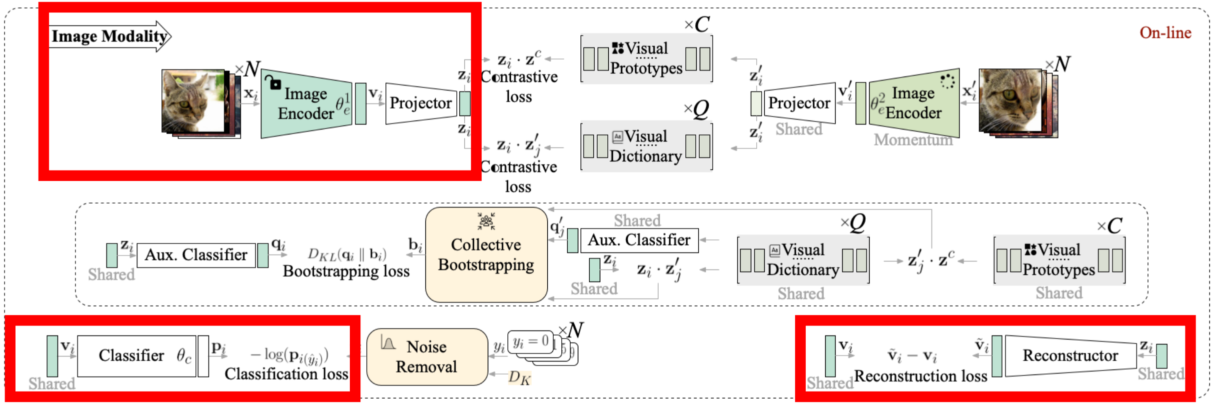


Vanilla

腾讯优图

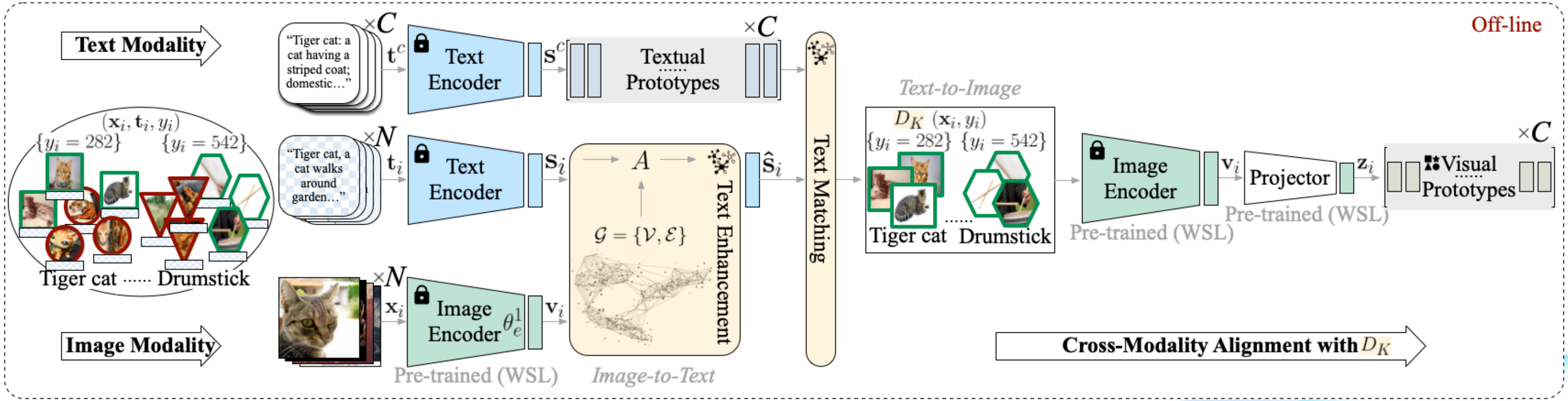
- Classification
- Projection and Reconstruction

$$\mathcal{L}_i^{\text{cls}} = -\log(\mathbf{p}_{i(y_i)}),\ \mathcal{L}_i^{\text{prj}} = \|\tilde{\mathbf{v}}_i - \mathbf{v}_i\|_2^2 - \log(\mathbf{q}_{i(y_i)}).$$



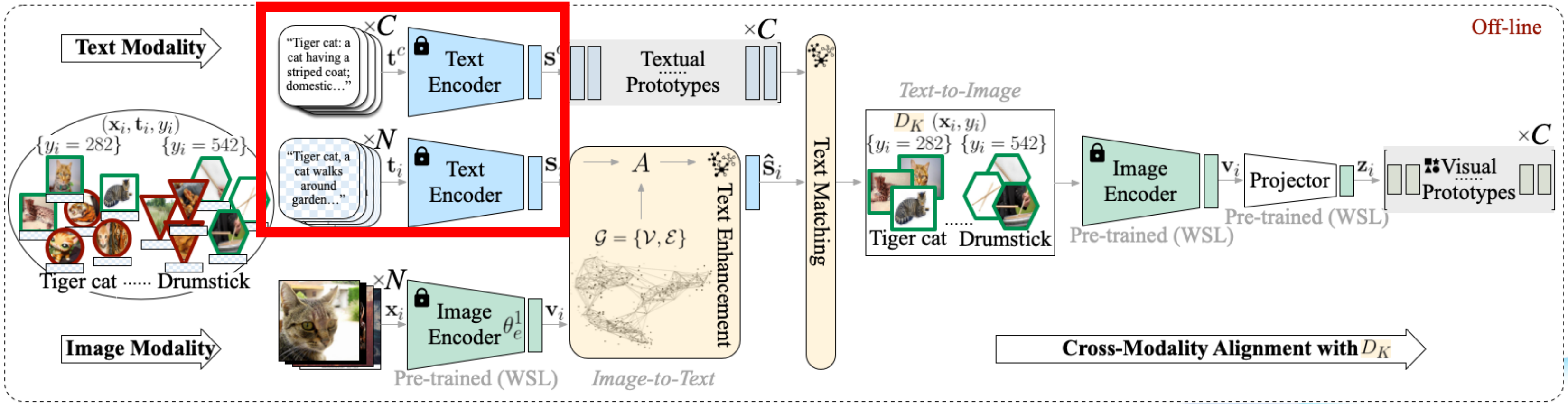
Cross-modality Alignment

- Text Encoding
- Text Enhancement
  - Visual Guidance from Neighbors
  - Reranking by k-reciprocal NNs
- Textual Prototypes
- Text Matching
- Visual Prototypes
- Noise Removal



Cross-modality Alignment

- **Text Encoding**
- Text Enhancement
  - Visual Guidance from Neighbors
  - Reranking by k-reciprocal NNs
- Textual Prototypes
- Text Matching
- Visual Prototypes
- Noise Removal



Cross-modality Alignment

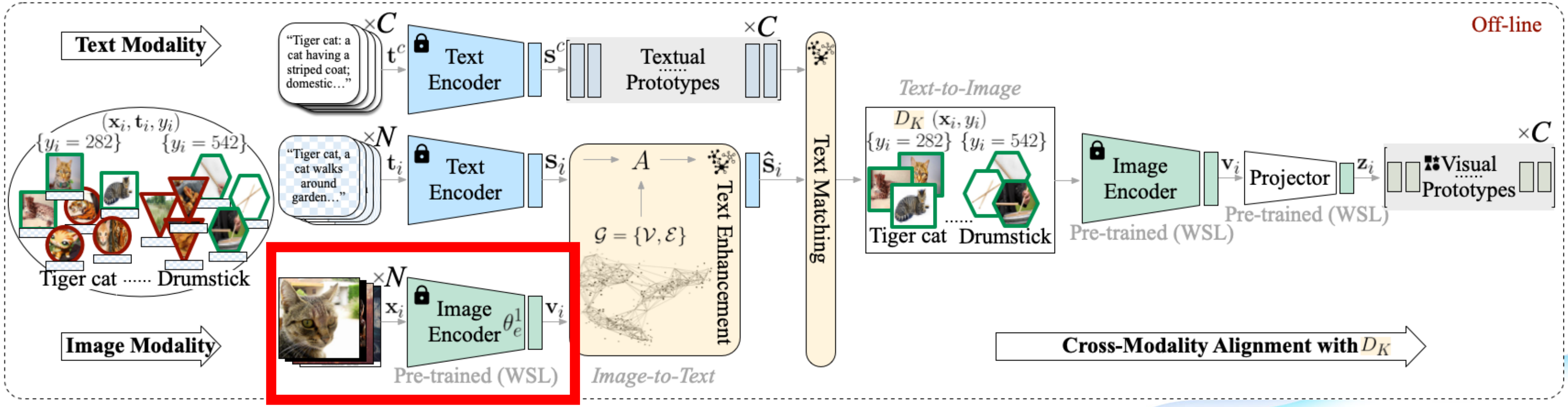
腾讯优图

- Text Encoding
- Text Enhancement
  - Visual Guidance from Neighbors
  - Reranking by k-reciprocal NNs
- Textual Prototypes
- Text Matching
- Visual Prototypes
- Noise Removal

$$A_{ij} = \begin{cases} 1 - d(\mathbf{v}_i, \mathbf{v}_j) & \text{, if } \mathbf{x}_i \in \mathcal{R}(\mathbf{x}_j, k) \text{ or } \mathbf{x}_j \in \mathcal{R}(\mathbf{x}_i, k), \\ 0 & \text{, otherwise,} \end{cases}$$

$$\mathcal{N}(\mathbf{x}_i, k) \text{ and } \mathcal{R}(\mathbf{x}_i, k) = \{\mathbf{x}_j | \mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i, k) \wedge \mathbf{x}_i \in \mathcal{N}(\mathbf{x}_j, k)\}$$

$$d(\mathbf{v}_i, \mathbf{v}_j) = 1 - \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}.$$



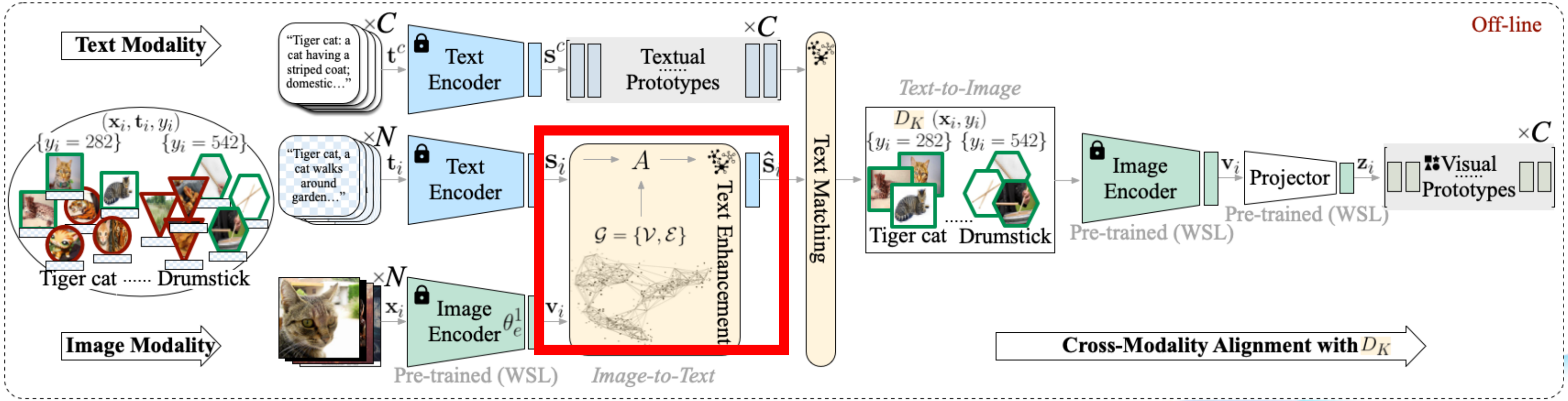
Cross-modality Alignment

- Text Encoding
- Text Enhancement
  - Visual Guidance from Neighbors
  - Reranking by k-reciprocal NNs
- Textual Prototypes
- Text Matching
- Visual Prototypes
- Noise Removal

$$d_J(\mathbf{v}_i,\mathbf{v}_j)=1-\frac{\sum_{k=1}^{N}\min(V_{\mathbf{v}_i,\mathbf{v}_k},V_{\mathbf{v}_j,\mathbf{v}_k})}{\sum_{k=1}^{N}\max(V_{\mathbf{v}_i,\mathbf{v}_k},V_{\mathbf{v}_j,\mathbf{v}_k})},\ V_{\mathbf{v}_i,\mathbf{v}_j}=\begin{cases}\exp(-d(\mathbf{v}_i,\mathbf{v}_j)) & ,\text{if}\,\mathbf{x}_j\in\mathcal{R}(\mathbf{x}_i,k)\\ 0 & ,\text{otherwise.}\end{cases}$$

$$d^*(\mathbf{v}_i,\mathbf{v}_j)=\frac{1}{2}(d(\mathbf{v}_i,\mathbf{v}_j)+d_J(\mathbf{v}_i,\mathbf{v}_j))$$

$$\mathbf{S}=(\mathbf{s}_1,\mathbf{s}_2,...,\mathbf{s}_N)\in\mathbb{R}^{N\times d_t}\ \hat{\mathbf{S}}=\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\mathbf{S},\tilde{A}=A+I_N,\tilde{D}_{ii}=\sum_j\tilde{A}_{ij}$$



Cross-modality Alignment

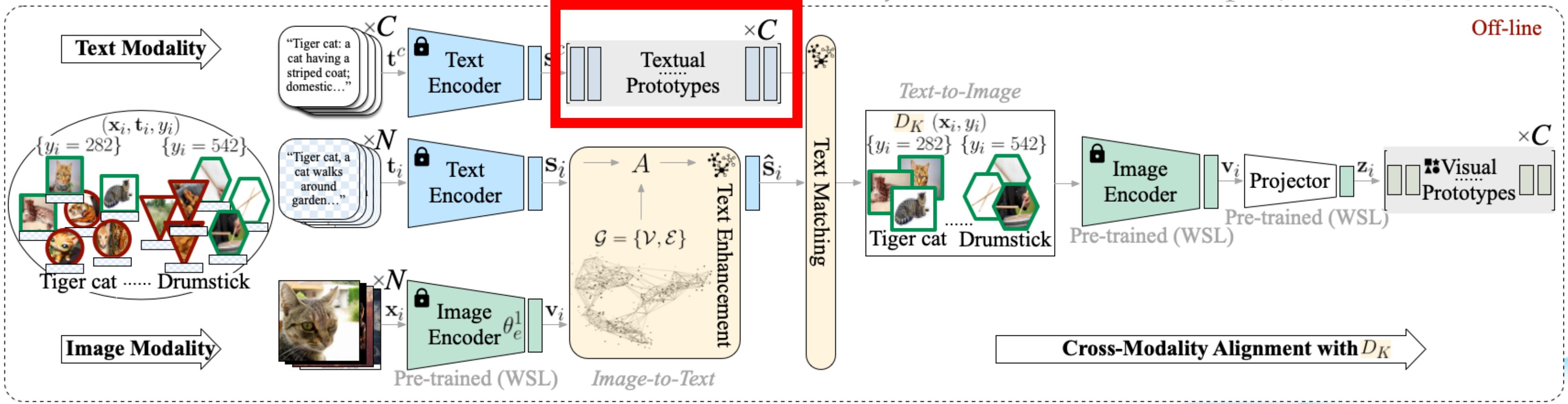
- Text Encoding
- Text Enhancement
  - Visual Guidance from Neighbors
  - Reranking by k-reciprocal NNs
- **Textual Prototypes**
- Text Matching
- Visual Prototypes
- Noise Removal

Keyword *tiger cat* (definition by **WordNet**)

+: *a cat having a striped coat; domestic_cat, house_cat, felis_domesticus, felis_catus: any domesticated member of the genus Felis*

−: *medium-sized wildcat in Central South America*

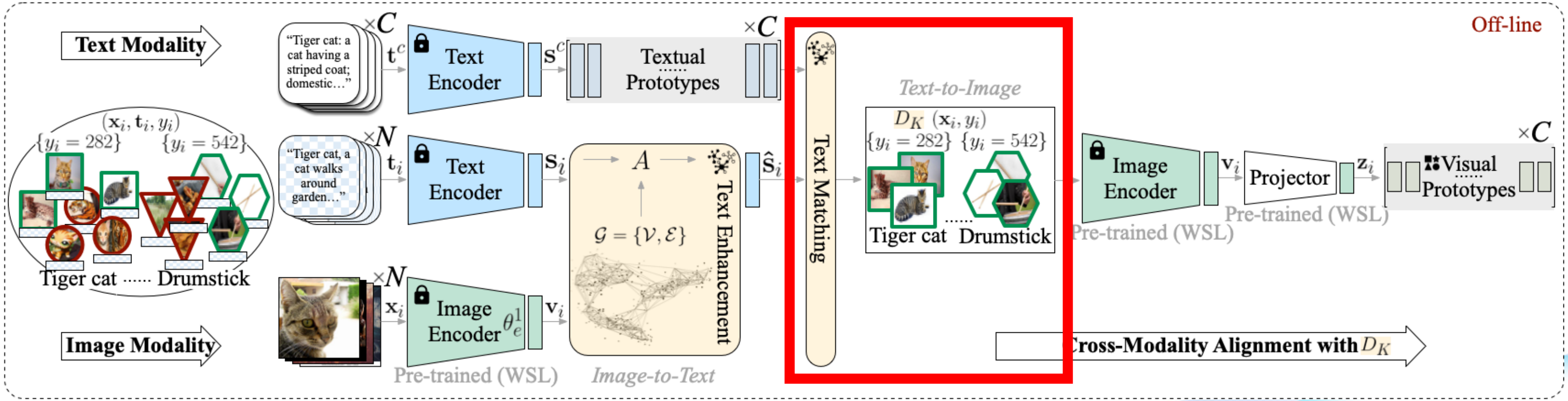
−: *large feline of forests in most of Asia having a tawny coat with black stripes; endangered*



Cross-modality Alignment

- Text Encoding
- Text Enhancement
    - Visual Guidance from Neighbors
    - Reranking by k-reciprocal NNs
- Textual Prototypes
- **Text Matching**
- Visual Prototypes
- Noise Removal

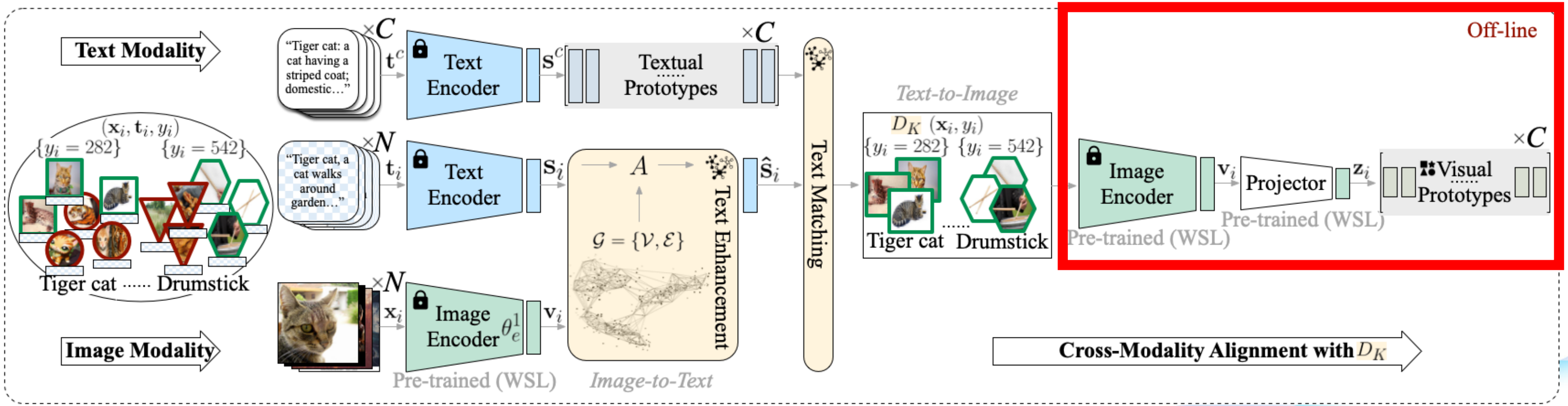
$$D_K = D_K^1 \cup D_K^2 \cup \ldots \cup D_K^C,\ D_K^c = \{(\mathbf{x}_i, \mathbf{t}_i, y_i) | (y_i = c) \wedge (d^*(\hat{\mathbf{s}}_i, \mathbf{s}^c) \leq \sigma_K^c)\},$$



Cross-modality Alignment

- Text Encoding
- Text Enhancement
    - Visual Guidance from Neighbors
    - Reranking by k-reciprocal NNs
- Textual Prototypes
- Text Matching
- **Visual Prototypes**
- Noise Removal

$$\hat{\mathbf{z}}^c = \frac{1}{K}\sum_{\mathbf{x}_i \in D_K^c} \mathbf{z}_i, \mathbf{z}^c = \frac{\hat{\mathbf{z}}^c}{\|\hat{\mathbf{z}}^c\|_2}$$

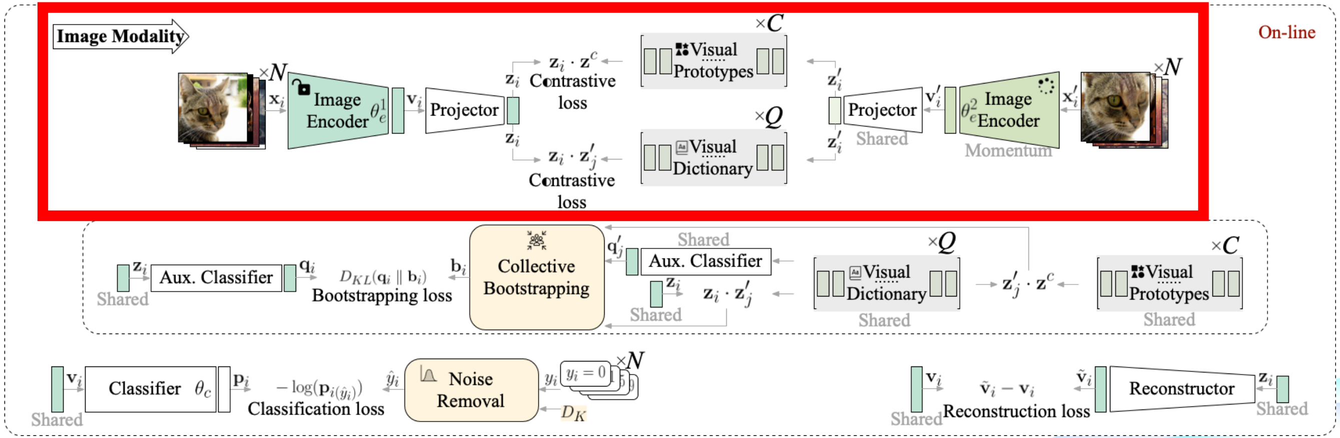


Cross-modality Alignment

- Text Encoding
- Text Enhancement
  - Visual Guidance from Neighbors
  - Reranking by k-reciprocal NNs
- Textual Prototypes
- Text Matching
- **Visual Prototypes**
- Noise Removal

$$\mathcal{L}_i^{\text{pro}} = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}^{y_i}/\tau)}{\sum_{c=1}^{C} \exp(\mathbf{z}_i \cdot \mathbf{z}^c/\tau)}, \quad \mathcal{L}_i^{\text{ins}} = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_i'/\tau)}{\sum_{j=1}^{Q} \exp(\mathbf{z}_i \cdot \mathbf{z}_j'/\tau)},$$

$$\hat{\mathbf{z}}^c = m_p \mathbf{z}^c + (1-m_p)\mathbf{z}_i, \ \mathbf{z}^c = \frac{\hat{\mathbf{z}}^c}{\|\hat{\mathbf{z}}^c\|_2}$$

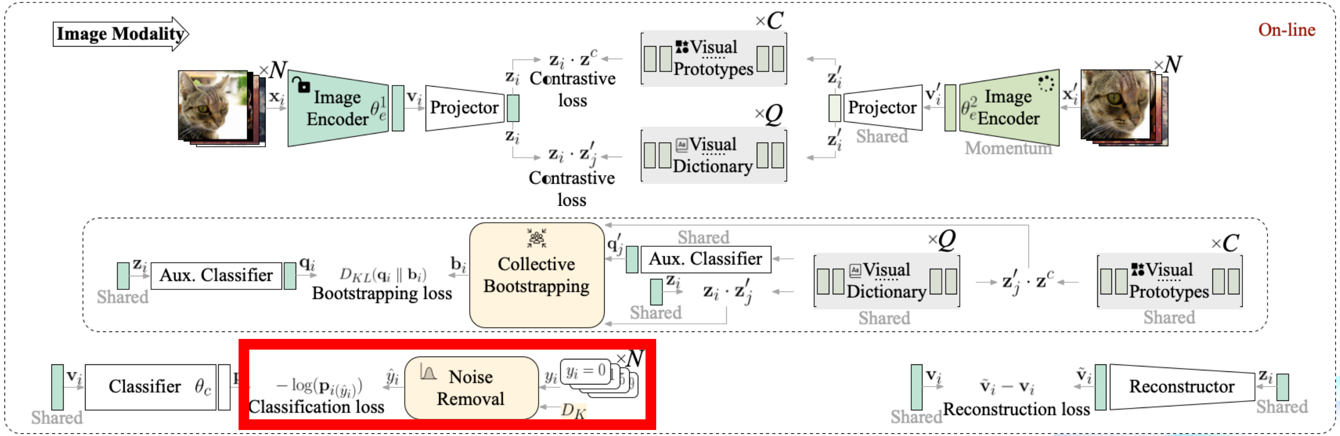


Cross-modality Alignment

- Text Encoding
- Text Enhancement
  - Visual Guidance from Neighbors
  - Reranking by k-reciprocal NNs
- Textual Prototypes
- Text Matching
- Visual Prototypes
- **Noise Removal**

$$\mathbf{o}_i = \alpha \mathbf{p}_i + (1-\alpha)\mathbf{r}_i,\ \ \mathbf{r}_{i(k)} = \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}^k/\tau)}{\sum_{c=1}^{C}\exp(\mathbf{z}_i \cdot \mathbf{z}^c/\tau)},$$

$$\hat{y}_i = \begin{cases} y_i & \text{if } \mathbf{x}_i \in D_K, \\ \arg\max_c \mathbf{o}_{i(c)} & \text{else if} \max_c \mathbf{o}_{i(c)} > \gamma, \\ y_i & \text{else if } \mathbf{o}_{i(y_i)} > 1/C, \\ \text{Null } (OOD) & \text{otherwise.} \end{cases}$$



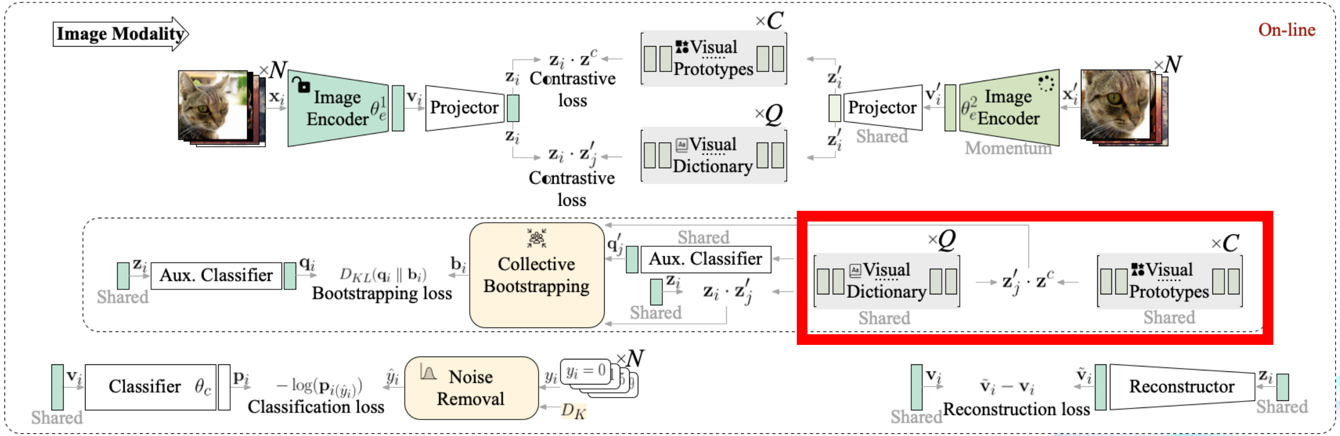
Collective Bootstrapping

- **Pseudo-label References**
- Query-Key Dictionary Look-up
- Bootstrapping

Self-prediction from the auxiliary classifier $-\log(\mathbf{q}_{i(y_i)})$.

Visual similarity with prototypes $\mathbf{r'}_{j(k)} = \frac{\exp(\mathbf{z}'_j \cdot \mathbf{z}^k/\tau)}{\sum_{c=1}^{C} \exp(\mathbf{z}'_j \cdot \mathbf{z}^c/\tau)}$

Weighted Pseudo-labels $(\alpha\mathbf{q}'_j + (1-\alpha)\mathbf{r}'_j)$.

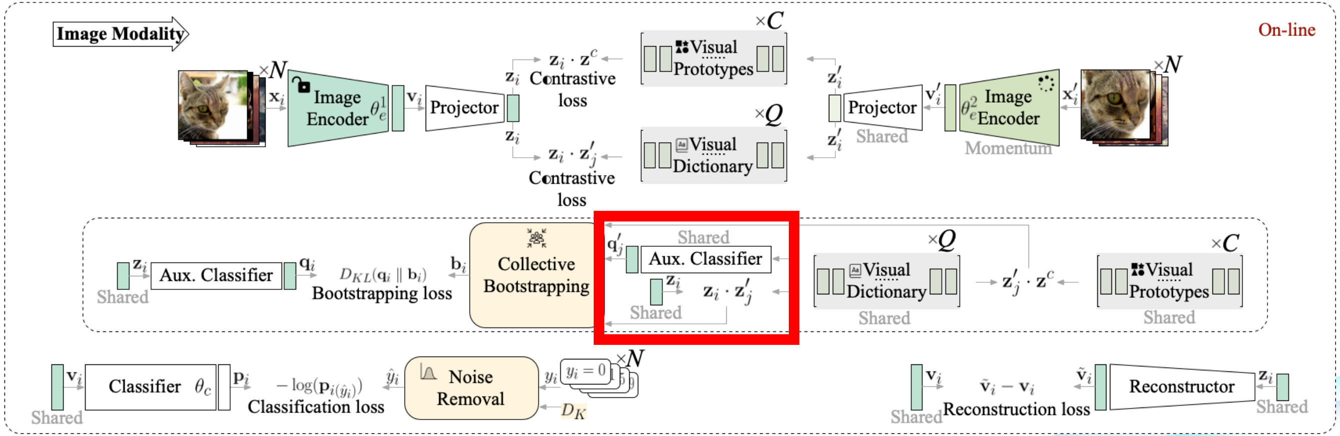


Collective Bootstrapping

- Pseudo-label References
- **Query-Key Dictionary Look-up**
- Bootstrapping

Similarity with NNs
in the visual dictionary

$$w_{ij} = \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}'_j/\tau)}{\sum_{j=1}^{Q} \exp(\mathbf{z}_i \cdot \mathbf{z}'_j/\tau)},$$



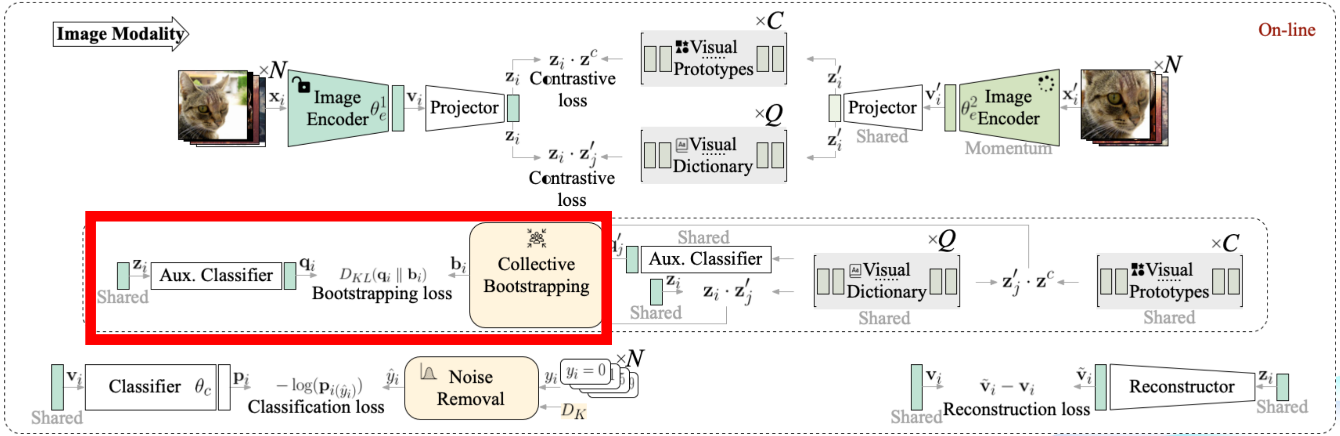
Collective Bootstrapping

- Pseudo-label References
- Query-Key Dictionary Look-up
- **Bootstrapping**

Bootstrapping targets

$$\mathbf{b}_i = \sum_{j=1}^{Q} w_{ij}(\alpha \mathbf{q}'_j + (1-\alpha)\mathbf{r}'_j),$$

$$\mathcal{L}_i^{\text{bts}} = D_{KL}(\mathbf{q}_i \parallel \mathbf{b}_i) = \sum_{c=1}^{C} \mathbf{q}_{i(c)} \log \frac{\mathbf{q}_{i(c)}}{\mathbf{b}_{i(c)}}.$$



Results

- Comparison with SOTA methods
    - Performance on single-label datasets
    - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- Ablation Study
    - Text Encoding and Enhancement
    - Reference Provider
    - λ^{bts} and Top-K
    - Threshold γ
    - Update Frequency of Prototypes
    - Noise Removal Policy
- Qualitative Results

Results

腾讯优图

- Comparison with SOTA methods
  - Performance on single-label datasets
  - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- Ablation Study
  - Text Encoding and Enhancement
  - Reference Provider
  - λ^{bts} and Top-K
  - Threshold γ
  - Update Frequency of Prototypes
  - Noise Removal Policy
- Qualitative Results

Table 1: Results on WebVision1k and Google500. Best/2nd best are marked bold/underlined.

Method	Back-bone	WebVision1k Top1	WebVision1k Top5	ImageNet1k Top1	ImageNet1k Top5	Google500 Top1	Google500 Top5	ImageNet500 Top1	ImageNet500 Top5
MentorNet [17]	IRV2 [90]	72.6	88.9	64.2	84.8	–	–	–	–
Curriculum [16]	IV2 [91]	72.1	89.1	64.8	84.9	–	–	–	–
Multimodal [92]	IV3 [93]	73.2	89.7	–	–	–	–	–	–
Vanilla [22]	R50D [94]	75.0	89.2	67.2	84.0	75.4	88.6	68.8	84.6
SCC [22]	R50D	<u>75.3</u>	89.3	67.9	84.7	**76.4**	<u>89.6</u>	<u>69.7</u>	<u>85.3</u>
Vanilla† [78]	R50	74.2	89.8	<u>68.2</u>	86.2	66.9	82.6	61.5	78.8
CoTeach [20; 78]	R50	–	–	–	–	67.6	84.0	62.1	80.9
VSGraph† [78]	R50	**75.4**	90.1	**69.4**	<u>87.2</u>	68.1	84.4	63.1	81.4
Vanilla [28]	R50	72.4	89.0	65.7	85.1	–	–	–	–
SOMNet [8]	R50	72.2	89.5	65.0	85.1	–	–	–	–
Curriculum [16]	R50	70.7	88.6	62.7	83.4	–	–	–	–
CleanNet [18]	R50	70.3	87.7	63.4	84.5	–	–	–	–
SINet [24]	R50	73.8	**90.6**	66.8	85.9	–	–	–	–
NCR [58]	R50	73.9	–	–	–	–	–	–	–
NCR† [58]	R50	75.7	–	–	–	–	–	–	–
MILe [26]	R50	75.2	90.3	67.1	85.6	–	–	–	–
MoPro [28]	R50	73.9	90.0	67.8	87.0	–	–	–	–
Vanilla (ours)	R50	72.6	89.7	67.0	86.8	69.9	86.5	64.5	83.1
CAPro (ours)	R50	74.2	<u>90.5</u>	68.0	**87.2**	<u>76.0</u>	**91.3**	**72.0**	**89.2**

† Results on WebVision1k are under optimized training settings with batch size of 1024.

Results

腾讯优图

- Comparison with SOTA methods
  - Performance on single-label datasets
  - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- Ablation Study
  - Text Encoding and Enhancement
  - Reference Provider
  - λ^{bts} and Top-K
  - Threshold γ
  - Update Frequency of Prototypes
  - Noise Removal Policy
- Qualitative Results

Table 2: Results on NUS-WIDE (Web).

Method	Back-bone	NUS-WIDE C-F1	NUS-WIDE O-F1	NUS-WIDE mAP
Vanilla [78]	R50	37.5	39.6	43.9
VSGraph [78]	R50	38.6	40.2	44.8
MCPL [95]	R101	22.5	17.2	47.4
Vanilla (ours)	R50	37.8	42.4	38.3
CAPro (ours)	R50	**39.3**	**45.4**	**48.0**

Results

- Comparison with SOTA methods
  - Performance on single-label datasets
  - Performance on multi-label datasets
- **Discussion on Open-Set Recognition**
- Ablation Study
  - Text Encoding and Enhancement
  - Reference Provider
  - λ^{bts} and Top-K
  - Threshold γ
  - Update Frequency of Prototypes
  - Noise Removal Policy
- Qualitative Results

Table 3: Results on open-set recognition.

Method	WebVision C-F1	ImageNet C-F1
Vanilla [78]	50.5	46.4
CoTeach [20; 78]	51.0	47.7
VSGraph [78]	57.2	52.8
Vanilla (ours)	54.6	48.3
CAPro (ours)	**62.2**	**57.8**

腾讯优图

Results

- Comparison with SOTA methods
  - Performance on single-label datasets
  - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- **Ablation Study**
  - **Text Encoding and Enhancement**
  - Reference Provider
  - λ^{bts} and Top-K
  - Threshold γ
  - Update Frequency of Prototypes
  - Noise Removal Policy
- Qualitative Results

Table 4: Ablation study on text encoding, enhancement, and reference provider.

Text Encoding	Text Enhancement	Reference Provider	Google500 Top1	Google500 Top5	ImageNet500 Top1	ImageNet500 Top5	NUS-WIDE C-F1	NUS-WIDE O-F1	NUS-WIDE mAP
×	×	×	71.5	87.8	66.5	84.6	37.2	42.4	46.2
MiniLM	VSGraph [78]	×	72.0	88.0	66.9	85.4	39.2	44.4	46.8
MiniLM	✓ (ours)	×	75.5	91.0	71.5	88.8	39.3	44.9	47.4
XLNet	VSGraph [78]	×	71.6	87.8	66.8	84.8	38.6	43.4	47.6
XLNet	✓ (ours)	×	75.4	91.0	71.5	88.8	39.3	45.1	47.5
GPT-Neo	VSGraph [78]	×	72.0	88.0	67.2	85.3	39.2	45.0	47.4
GPT-Neo	✓ (ours)	×	75.7	91.1	71.6	88.8	39.2	45.1	47.6
MiniLM	✓ (ours)	Mix-up (MU) [99]	75.7	90.9	71.4	88.6	38.7	45.3	47.2
MiniLM	✓ (ours)	Bootstrap [33]	75.5	90.8	71.3	88.4	38.1	43.2	46.0
MiniLM	✓ (ours)	Label smooth [100]	75.4	90.8	71.2	88.4	36.9	42.1	46.8
MiniLM	✓ (ours)	SCC [22]	73.8	89.9	70.2	88.0	35.6	41.3	45.0
MiniLM	✓ (ours)	NCR [58]	75.5	91.1	71.5	88.8	37.6	43.4	46.8
MiniLM	✓ (ours)	✓ CB (ours)	76.0	91.3	72.0	89.2	39.3	45.4	48.0
MiniLM	✓ (ours)	✓ CB (ours) + MU	76.5	91.1	71.9	88.8	40.4	**46.7**	49.9
GPT-Neo	✓ (ours)	✓ CB (ours)	76.1	**91.4**	**72.1**	**89.4**	39.3	44.9	47.7
GPT-Neo	✓ (ours)	✓ CB (ours) + MU	**76.5**	91.2	72.0	88.8	**40.7**	45.2	**50.0**

Results

- Comparison with SOTA methods
  - Performance on single-label datasets
  - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- Ablation Study
  - Text Encoding and Enhancement
  - Reference Provider
  - λ^{bts} and Top-K
  - Threshold γ
  - Update Frequency of Prototypes
  - Noise Removal Policy
- Qualitative Results

Table 4: Ablation study on text encoding, enhancement, and reference provider.

Text Encoding	Text Enhan-cement	Reference Provider	Google500 Top1	Google500 Top5	ImageNet500 Top1	ImageNet500 Top5	NUS-WIDE C-F1	NUS-WIDE O-F1	NUS-WIDE mAP
×	×	×	71.5	87.8	66.5	84.6	37.2	42.4	46.2
MiniLM	VSGraph [78]	×	72.0	88.0	66.9	85.4	39.2	44.4	46.8
MiniLM	✓ (ours)	×	75.5	91.0	71.5	88.8	39.3	44.9	47.4
XLNet	VSGraph [78]	×	71.6	87.8	66.8	84.8	38.6	43.4	47.6
XLNet	✓ (ours)	×	75.4	91.0	71.5	88.8	39.3	45.1	47.5
GPT-Neo	VSGraph [78]	×	72.0	88.0	67.2	85.3	39.2	45.0	47.4
GPT-Neo	✓ (ours)	×	75.7	91.1	71.6	88.8	39.2	45.1	47.6
MiniLM	✓ (ours)	Mix-up (MU) [99]	75.7	90.9	71.4	88.6	38.7	45.3	47.2
MiniLM	✓ (ours)	Bootstrap [33]	75.5	90.8	71.3	88.4	38.1	43.2	46.0
MiniLM	✓ (ours)	Label smooth [100]	75.4	90.8	71.2	88.4	36.9	42.1	46.8
MiniLM	✓ (ours)	SCC [22]	73.8	89.9	70.2	88.0	35.6	41.3	45.0
MiniLM	✓ (ours)	NCR [58]	75.5	91.1	71.5	88.8	37.6	43.4	46.8
MiniLM	✓ (ours)	✓ CB (ours)	76.0	91.3	72.0	89.2	39.3	45.4	48.0
MiniLM	✓ (ours)	✓ CB (ours) + MU	76.5	91.1	71.9	88.8	40.4	**46.7**	49.9
GPT-Neo	✓ (ours)	✓ CB (ours)	76.1	**91.4**	**72.1**	**89.4**	39.3	44.9	47.7
GPT-Neo	✓ (ours)	✓ CB (ours) + MU	**76.5**	91.2	72.0	88.8	**40.7**	45.2	**50.0**

Results

- Comparison with SOTA methods
  - Performance on single-label datasets
  - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- **Ablation Study**
  - Text Encoding and Enhancement
  - Reference Provider
  - **λ^{bts} and Top-K**
  - Threshold γ
  - Update Frequency of Prototypes
  - Noise Removal Policy
- Qualitative Results

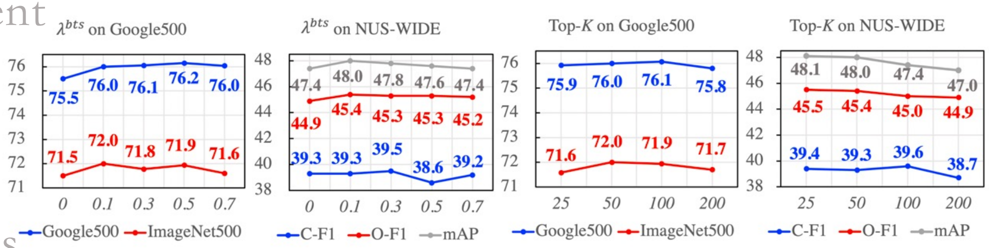


Figure 4: Impact of hyper-parameters λ^{bts} and top-K on CAPro.

Results

腾讯优图

- Comparison with SOTA methods
  - Performance on single-label datasets
  - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- **Ablation Study**
  - Text Encoding and Enhancement
  - Reference Provider
  - λ^{bts} and Top-K
  - **Threshold γ**
  - Update Frequency of Prototypes
  - Noise Removal Policy
- Qualitative Results

Table 7: Effect of γ on CAPro without collective bootstrapping.

γ	Reference Provider	Google500		ImageNet500		NUS-WIDE		
		Top1	Top5	Top1	Top5	C-F1	O-F1	mAP
0.6	×	72.0	88.0	66.9	85.4	8.3	9.1	6.9
0.8	×	71.2	87.7	65.9	84.8	–	–	–
0.9	×	–	–	–	–	39.2	44.4	46.8

Results

- Comparison with SOTA methods
  - Performance on single-label datasets
  - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- **Ablation Study**
  - Text Encoding and Enhancement
  - Reference Provider
  - λ^{bts} and Top-K
  - Threshold γ
  - **Update Frequency of Prototypes**
  - Noise Removal Policy
- Qualitative Results

Table 8: Effect of prototype update frequency on CAPro. By default, we update visual prototypes every epoch using high-quality examples in each mini-batch. For 0-epoch per update, we do not introduce additional high-quality web examples to polish prototypes, but only update them with the top-K matched semantically-correct examples with their latest visual embeddings.

# Epochs	Google500		ImageNet500		NUS-WIDE		
per update	Top1	Top5	Top1	Top5	C-F1	O-F1	mAP
0	75.5	91.1	71.6	88.8	39.2	44.4	47.2
1 (by default)	76.0	91.3	72.0	89.2	39.3	45.4	48.0
5	75.9	91.2	71.8	89.2	39.6	45.0	47.6
10	76.0	91.2	71.7	89.1	39.3	45.8	48.2

Results

- Comparison with SOTA methods
    - Performance on single-label datasets
    - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- **Ablation Study**
    - Text Encoding and Enhancement
    - Reference Provider
    - λ^{bts} and Top-K
    - Threshold γ
    - Update Frequency of Prototypes
    - **Noise Removal Policy**
- Qualitative Results

Table 9: Effect of noise removal policy on CAPro. We compare with MoPro to show the effectiveness of keeping labels of top-K matched semantically-correct examples unchanged.

Noise Removal	Google500		ImageNet500		NUS-WIDE		
policy	Top1	Top5	Top1	Top5	C-F1	O-F1	mAP
MoPro [28]	75.8	91.1	71.7	89.0	38.8	42.2	47.2
CAPro (ours)	76.0	91.3	72.0	89.2	39.3	45.4	48.0

Results

腾讯优图

- Comparison with SOTA methods
  - Performance on single-label datasets
  - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- Ablation Study
  - Text Encoding and Enhancement
  - Reference Provider
  - λ^{bts} and Top-K
  - Threshold γ
  - Update Frequency of Prototypes
  - Noise Removal Policy
- **Qualitative Results**

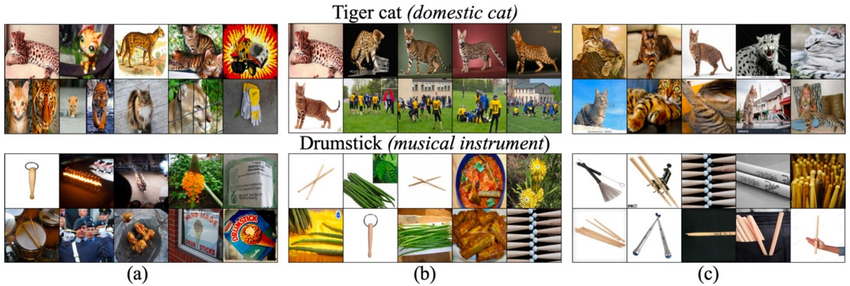


Figure 3: Top-matched WebVision1k instances are chosen: (a) without text enhancement, (b) with text enhancement in VSGraph [78], and (c) with our text enhancement.

Results

腾讯优图

- Comparison with SOTA methods
    - Performance on single-label datasets
    - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- Ablation Study
    - Text Encoding and Enhancement
    - Reference Provider
    - λ^{bts} and Top-K
    - Threshold γ
    - Update Frequency of Prototypes
    - Noise Removal Policy
- **Qualitative Results**

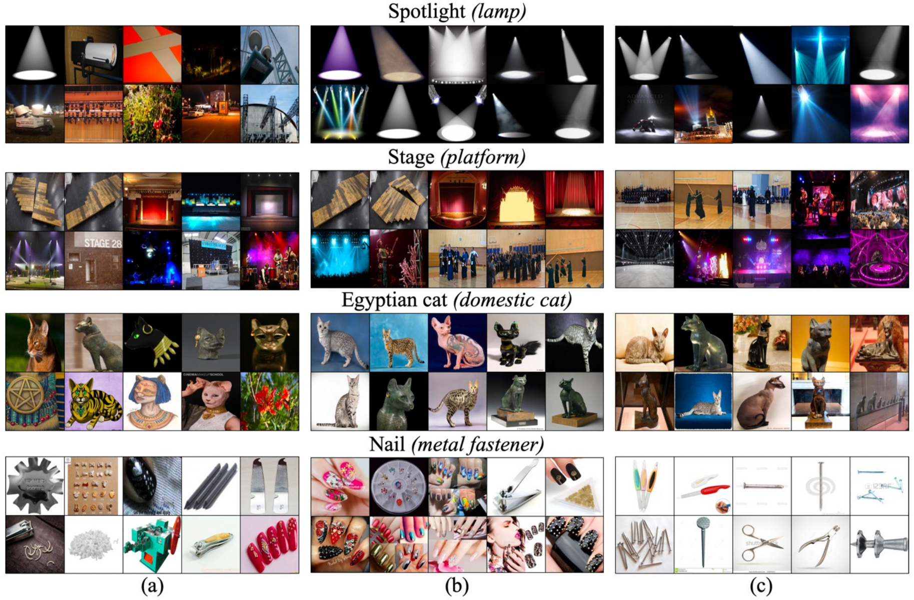


Figure 5: Top-matched WebVision1k instances are chosen: (a) without text enhancement, (b) with text enhancement in VSGraph [78], and (c) with our text enhancement.

Results

- Comparison with SOTA methods
  - Performance on single-label datasets
  - Performance on multi-label datasets
- Discussion on Open-Set Recognition
- Ablation Study
  - Text Encoding and Enhancement
  - Reference Provider
  - λ^{bts} and Top-K
  - Threshold γ
  - Update Frequency of Prototypes
  - Noise Removal Policy
- **Qualitative Results**

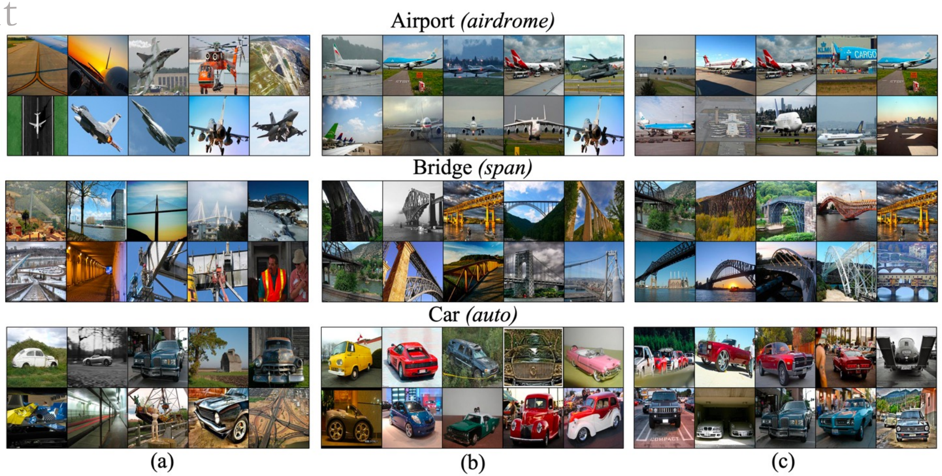


Figure 6: Top-matched NUS-WIDE (Web) instances are chosen: (a) without text enhancement, (b) with text enhancement in VSGraph [78], and (c) with our text enhancement.

腾讯优图

Thanks for your attention!