CLIP-driven Coarse-to-fine Semantic Guidance for Fine-grained Open-set Semi-supervised Learning

Xiaokun Li, Yaping Huang, Qingji Guan
Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing Jiaotong University

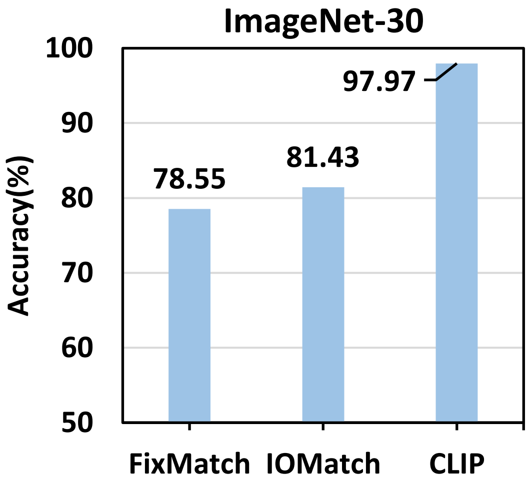
Problem Formulation

Fine-grained open-set semi-supervised learning (OSSL) investigates a practical scenario where unlabeled data may contain fine-grained out-of-distribution (OOD) samples. Due to the subtle visual differences among in-distribution (ID) samples, as well as between ID and OOD samples, it is extremely challenging to separate the ID and OOD samples.

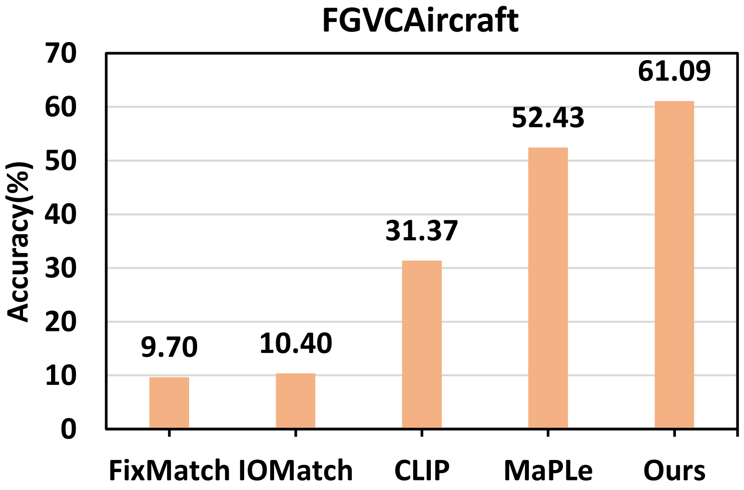
Problem Formulation

Comparison of OSSL methods on coarse-grained and fine-grained classification tasks.

(a) Coarse-grained OSSL



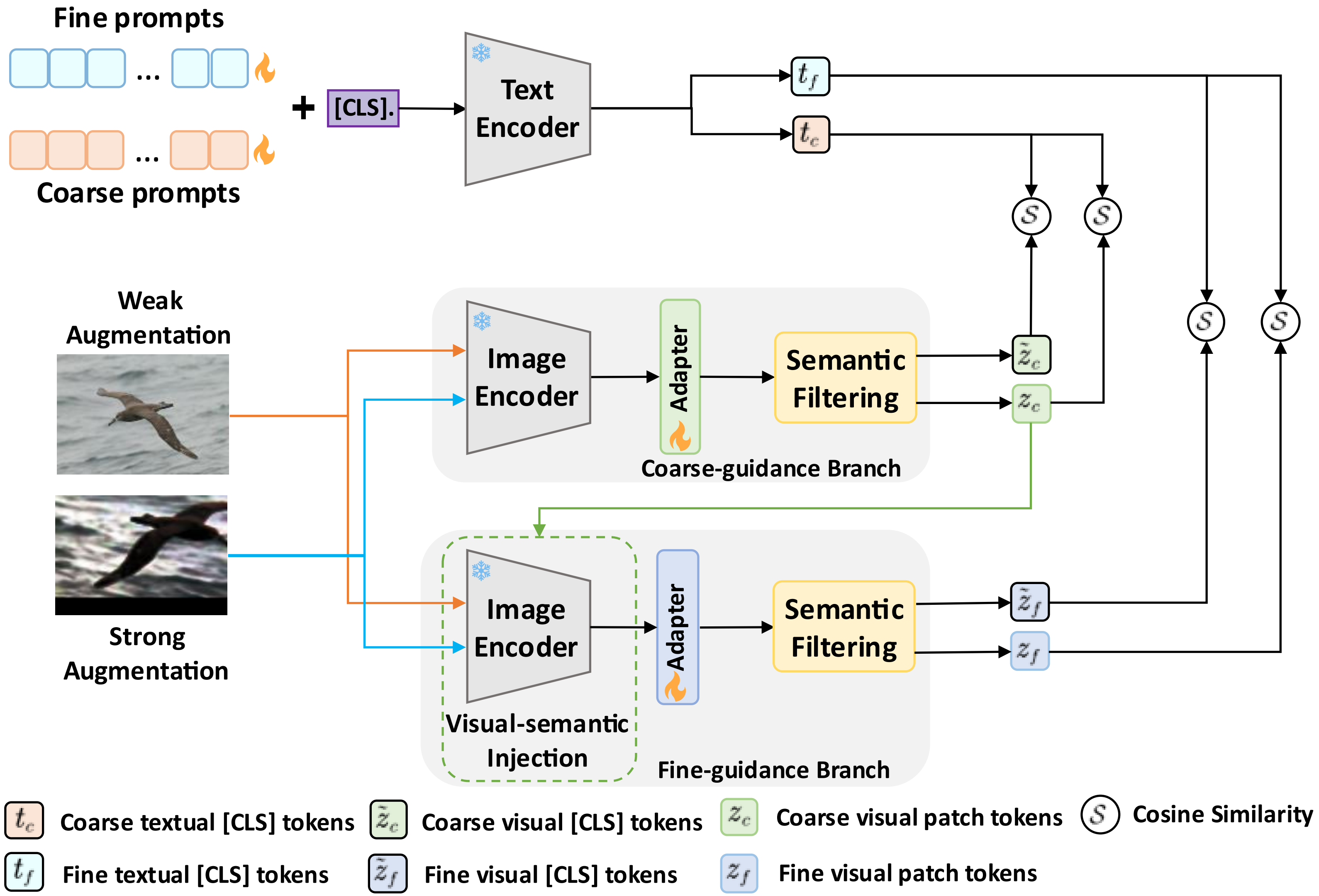
(b) Fine-grained OSSL



- ✓ On the coarse-grained ImageNet-30 dataset, some out-standing methods (e.g., FixMatch, IOMatch) achieve excellent performance. CLIP trained on the large-scale image-text pairs dataset achieves high performance as expected.
- ✓ On the fine-grained FGVCAircraft dataset, the generalization capabilities of CLIP remain effective but limited, as it tends to focus on general attributes while failing to capture the fine-grained details.

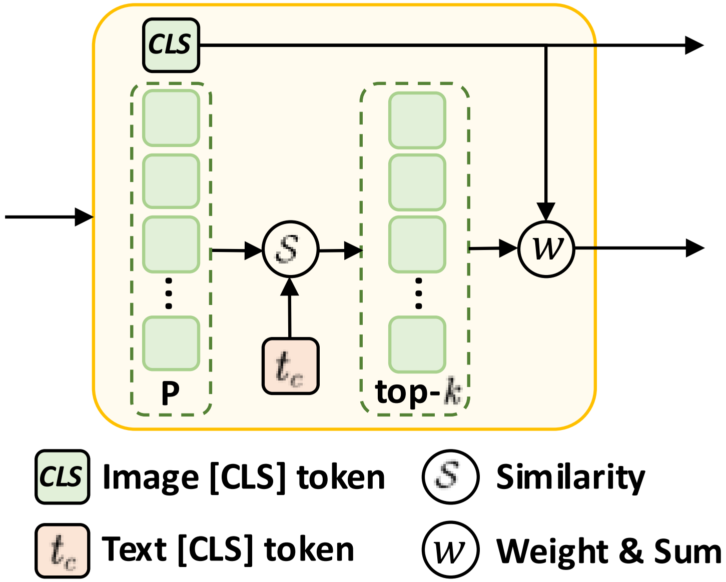
Method

Overview of the proposed CFSG-CLIP framework.



CFSG-CLIP is composed of a coarse-guidance branch and a fine-guidance branch based on the pre-trained CLIP model.

Semantic Filtering Module



Top-k patch-level visual features:

$$s_{c_i} = \mathrm{sim}(z_{c_i}, t_c^m),$$

$$\mathcal{K} = \{i \in P : \mathrm{rank}(s_{c_i}) \leq k\},$$

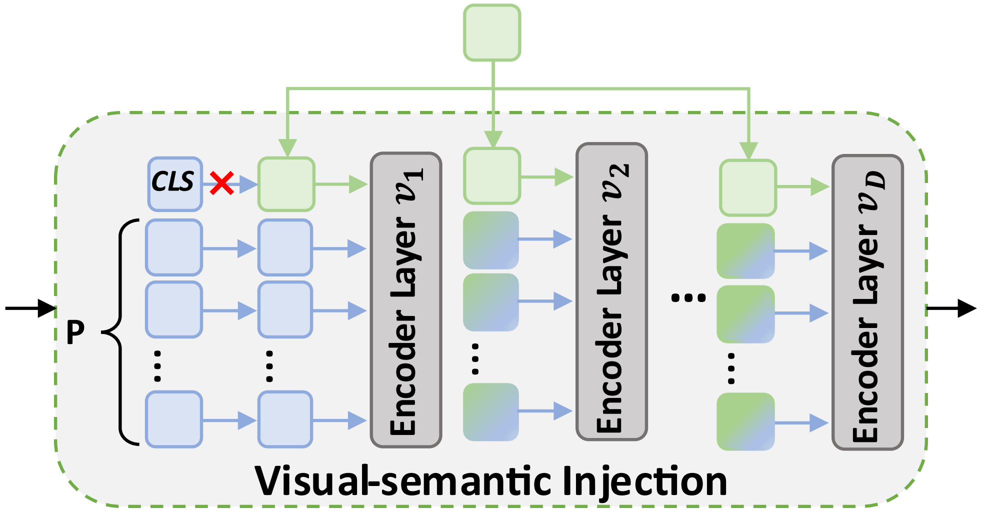
Image-level global feature weighting:

$$w_{c_i} = \frac{\exp(\mathrm{sim}(z_{c_i}, \tilde{z}_c))}{\sum \exp(\mathrm{sim}(z_{c_i}, \tilde{z}_c))}, i \in \mathcal{K}$$

$$z_c = \sum_{i \in \mathcal{K}} w_{c_i} z_{c_i}$$

To capture fine-grained details and maintain semantic consistency, we design a semantic filtering module that leverages both textual and visual features to obtain global and local visual features.

Semantic Filtering Module



Injecting the local visual features into transformer blocks:

$$[\_, E_j] = \mathcal{V}_j([\text{proj}(z_c), E_{j-1}]) \quad j = 1, 2, \ldots, D,$$

$$[\tilde{z}_{f_j}, E_j] = \mathcal{V}_j([\text{proj}(z_c), E_{j-1}]) \quad j = D+1, \ldots, T,$$

To guide the visual encoder to focus more on the fine-grained details of the image, we further design a visual semantic injection strategy in the fine-guidance branch.

Method

Dual-branch Training

Optimization of Coarse-guidance Branch:

$$\tilde{p}_c^l = \frac{\exp(\mathrm{sim}(\tilde{z}_c^l, t_c^m)/\tau)}{\sum_{m'} \exp(\mathrm{sim}(\tilde{z}_c^l, t_c^{m'})/\tau)},$$

$$p_c^l = \frac{\exp(\mathrm{sim}(z_c^l, t_c^m)/\tau)}{\sum_{m'} \exp(\mathrm{sim}(z_c^l, t_c^{m'})/\tau)},$$

$$L_c = H(y^l, \tilde{p}_c^l) + H(y^l, p_c^l) + \lambda_c(\mathcal{F}(x^u) H(\tilde{p}_c^{u_w}, \tilde{p}_c^{u_s}) \\ + \mathcal{F}(x^u) H(p_c^{u_w}, p_c^{u_s})),$$

Method

Dual-branch Training

Optimization of Fine-guidance Branch:

$$\tilde{p}_f^l = \frac{\exp(\mathrm{sim}(\tilde{z}_f^l, t_f^m)/\tau)}{\sum_{m'} \exp(\mathrm{sim}(\tilde{z}_f^l, t_f^{m'})/\tau)}$$

$$p_f^l = \frac{\exp(\mathrm{sim}(z_f^l, t_f^m)/\tau)}{\sum_{m'} \exp(\mathrm{sim}(z_f^l, t_f^{m'})/\tau)}$$

$$L_f = H(y^l, \tilde{p}_f^l) + H(y^l, p_f^l) + \lambda_f(\mathcal{F}(\tilde{u}_f)H(\tilde{p}_f^{u_w}, \tilde{p}_f^{u_s}) + \mathcal{F}(u_f)H(p_f^{u_w}, p_f^{u_s})),$$

Dual-branch Training

Optimization of Fine-guidance Branch:

$$\tilde{p}_f^l = \frac{\exp(\mathrm{sim}(\tilde{z}_f^l, t_f^m)/\tau)}{\sum_{m'} \exp(\mathrm{sim}(\tilde{z}_f^l, t_f^{m'})/\tau)}$$

$$p_f^l = \frac{\exp(\mathrm{sim}(z_f^l, t_f^m)/\tau)}{\sum_{m'} \exp(\mathrm{sim}(z_f^l, t_f^{m'})/\tau)}$$

$$\begin{aligned} L_f = H(y^l, \tilde{p}_f^l) + H(y^l, p_f^l) + \lambda_f(\mathcal{F}(\tilde{u}_f) H(\tilde{p}_f^{u_w}, \tilde{p}_f^{u_s}) \\ + \mathcal{F}(u_f) H(p_f^{u_w}, p_f^{u_s})), \end{aligned}$$

Experiments

Table 1. Classification accuracy (%) for CLIP-based methods on four fine-grained benchmark datasets with varying labeled set sizes under the fine-grained OSSL setting. The results are presented as the mean with standard deviation over three runs using different random seeds.

Method	Stanford Dogs		Stanford Cars		CUB-200-2011		FGVCAircraft	
	5	20	5	20	5	20	5	20
CLIP [27]	79.25±0.00	79.25±0.00	75.97±0.00	75.97±0.00	66.00±0.00	66.00±0.00	31.37±0.00	31.37±0.00
CLIP-LORA [41]	83.81±0.37	84.31±0.27	82.71±0.56	82.45±1.30	70.95±0.85	73.40±0.75	40.89±1.73	42.37±0.71
CLIP-Adapter [7]	82.91±0.25	86.02±0.27	84.31±0.02	87.13±0.28	80.03±0.29	84.77±1.02	47.77±0.90	55.79±0.73
CoOp [45]	83.01±0.26	85.68±0.37	85.45±0.31	87.64±0.46	80.10±0.29	85.40±0.37	45.39±0.96	55.43±0.30
LoCoOp [24]	83.08±0.25	86.26±0.11	84.10±0.72	87.83±0.66	79.27±0.45	85.63±0.54	45.53±1.36	54.67±1.59
PLOT [3]	84.46±0.07	87.11±0.09	86.28±0.30	88.59±0.45	81.43±0.66	87.20±0.14	49.59±0.37	58.25±0.93
MaPLe [14]	**85.64±0.15**	87.64±0.20	88.16±0.25	90.34±0.25	83.30±0.33	88.77±0.21	52.43±0.47	64.33±1.21
Ours	85.48±0.21	**89.42±0.16**	**90.38±0.09**	**93.08±0.08**	**84.73±0.17**	**91.75±0.24**	**61.09±0.27**	**73.56±0.58**

Table 2. Open-set classification balanced accuracy (%) for CLIP-based methods on four fine-grained benchmark datasets with varying labeled set sizes under the fine-grained OSSL setting.

Method	Stanford Dogs		Stanford Cars		CUB-200-2011		FGVCAircraft	
	5	20	5	20	5	20	5	20
CLIP [27]	77.17±0.00	77.17±0.00	75.70±0.00	75.70±0.00	64.10±0.00	64.10±0.00	31.08±0.00	31.08±0.00
CLIP-LORA [41]	82.34±0.57	82.67±0.36	82.10±0.56	81.03±1.24	70.52±1.34	72.20±0.45	40.10±1.69	41.57±0.69
CLIP-Adapter [7]	81.63±0.14	84.36±0.25	83.65±0.04	86.50±0.25	81.36±0.60	85.20±0.95	46.87±0.88	53.52±1.35
CoOp [45]	81.65±0.20	84.25±0.41	84.63±0.36	86.84±0.42	81.04±0.06	84.92±0.51	44.55±0.94	54.35±0.30
LoCoOp [24]	81.78±0.23	84.68±0.20	83.25±0.76	87.17±0.64	80.45±0.86	85.46±0.32	44.67±1.35	53.60±1.57
PLOT [3]	82.95±0.07	85.54±0.11	85.58±0.32	87.83±0.34	83.32±0.31	87.16±0.17	48.68±0.37	57.13±0.92
MaPLe [14]	**84.09±0.19**	86.02±0.25	87.43±0.27	89.48±0.25	84.72±0.53	88.66±0.60	51.79±0.43	63.12±1.19
Ours	84.02±0.15	**87.77±0.19**	**89.65±0.05**	**92.34±0.10**	**86.46±0.25**	**90.92±0.24**	**59.92±0.26**	**72.13±0.57**

Experiments

Table 3. Classification accuracy (%) on the Semi-Aves dataset is reported for two settings: unlabeled data with ID samples U_{in}, and unlabeled data with a mix of ID and OOD samples $U_{in} + U_{out}$.

Method	U_{in}	$U_{in} + U_{out}$
CLIP [27]	10.05±0.00	10.05±0.00
CLIP-Adapter [7]	50.16±1.55	50.10±0.44
CoOp [45]	47.67±0.97	47.93±0.76
LoCoOp [24]	48.04±0.96	47.30±1.30
PLOT [3]	53.68±0.35	54.72±0.96
MaPLe [14]	60.39±0.10	59.68±0.64
Ours	**65.63**±**0.27**	**63.69**±**0.74**

Experiments

Table 4. Ablation studies on the Stanford cars and FGVCAircraft datasets. The '$\mathcal{A}$' stands for adapter, 'SFM' denotes semantic filtering module, 'C' means coarse-guidance branch, 'F' means fine-guidance branch, and 'VSI' is visual-semantic injection strategy. The results are reported based on a single run with seed 1.

Method	Stanford Cars	FGVCAircraft
CoOp+$\mathcal{A}$	86.31	50.42
+SFM (C)	89.15	54.65
+SFM (C)+SFM (F)	90.03	58.01
+SFM (C)+SFM (F)+VSI	90.28	61.07

Table 5. Ablation studies for semantic filtering module. The results are reported based on a single run with seed 1.

Method	Stanford Cars	FGVCAircraft
CoOp+$\mathcal{A}$	86.31	50.42
Textual Filtering	88.65	53.57
Visual Weighting	89.15	54.65

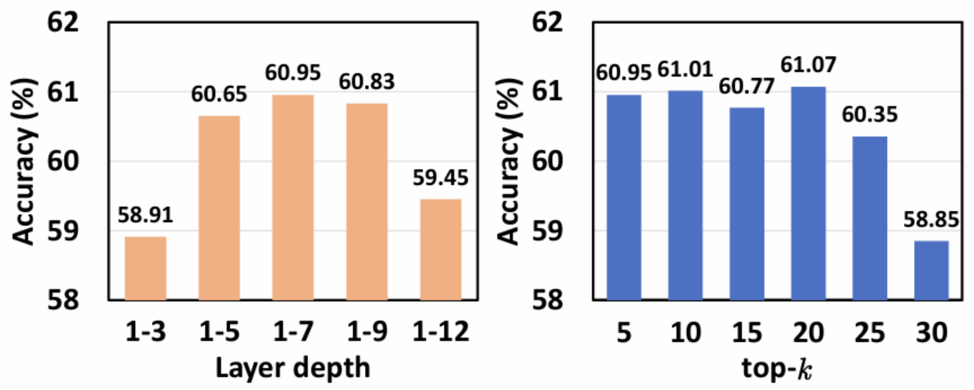


Figure 6. Ablation studies on injection depth (*left*) and top-k (*right*) on the FGVCAircraf datasets. The results are reported based on a single run with seed 1.

Table 6. Evaluation results for different operations of employing local visual features. The results are reported based on a single run with seed 1.

Method	Stanford Cars	FGVCAircraft
Concatenate	89.99	60.47
Replace	90.28	61.07

Experiments

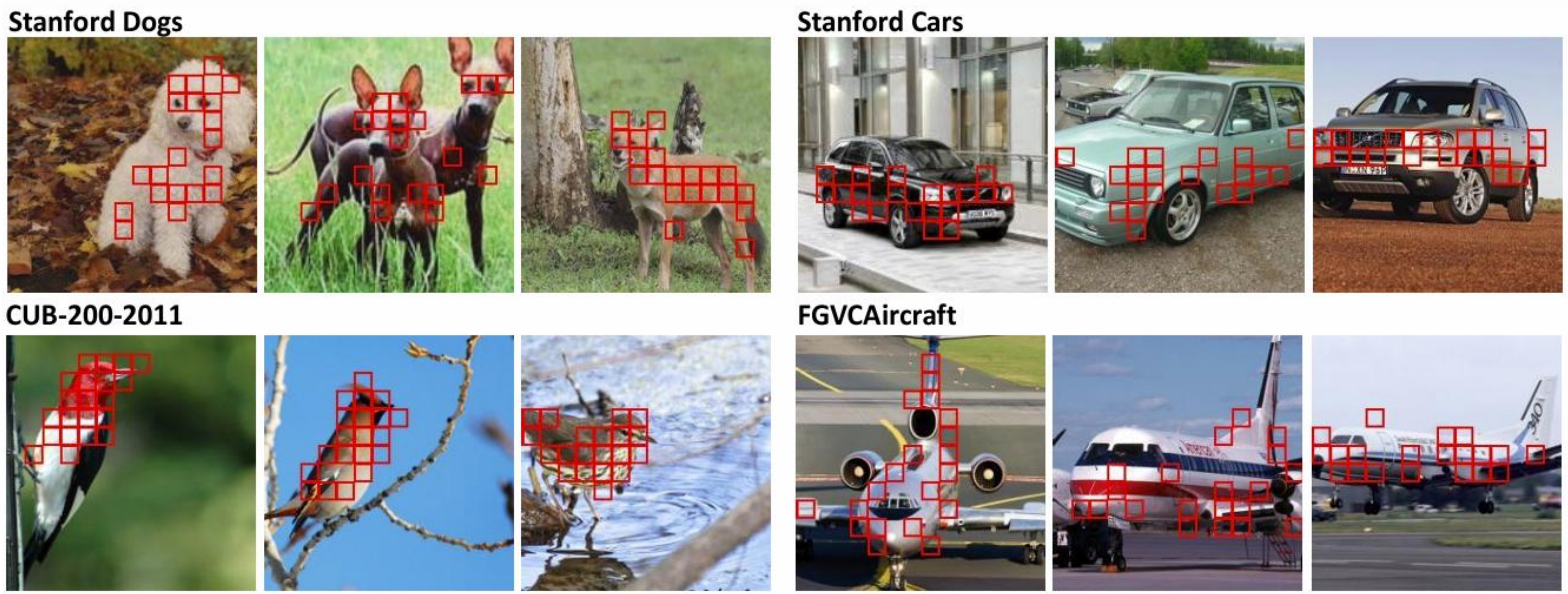


Figure 5. Visualization of patch-tokens extracted by semantic filtering module. We find that the semantic filtering module can correctly extract local visual regions on different fine-grained datasets.

Thank you!