

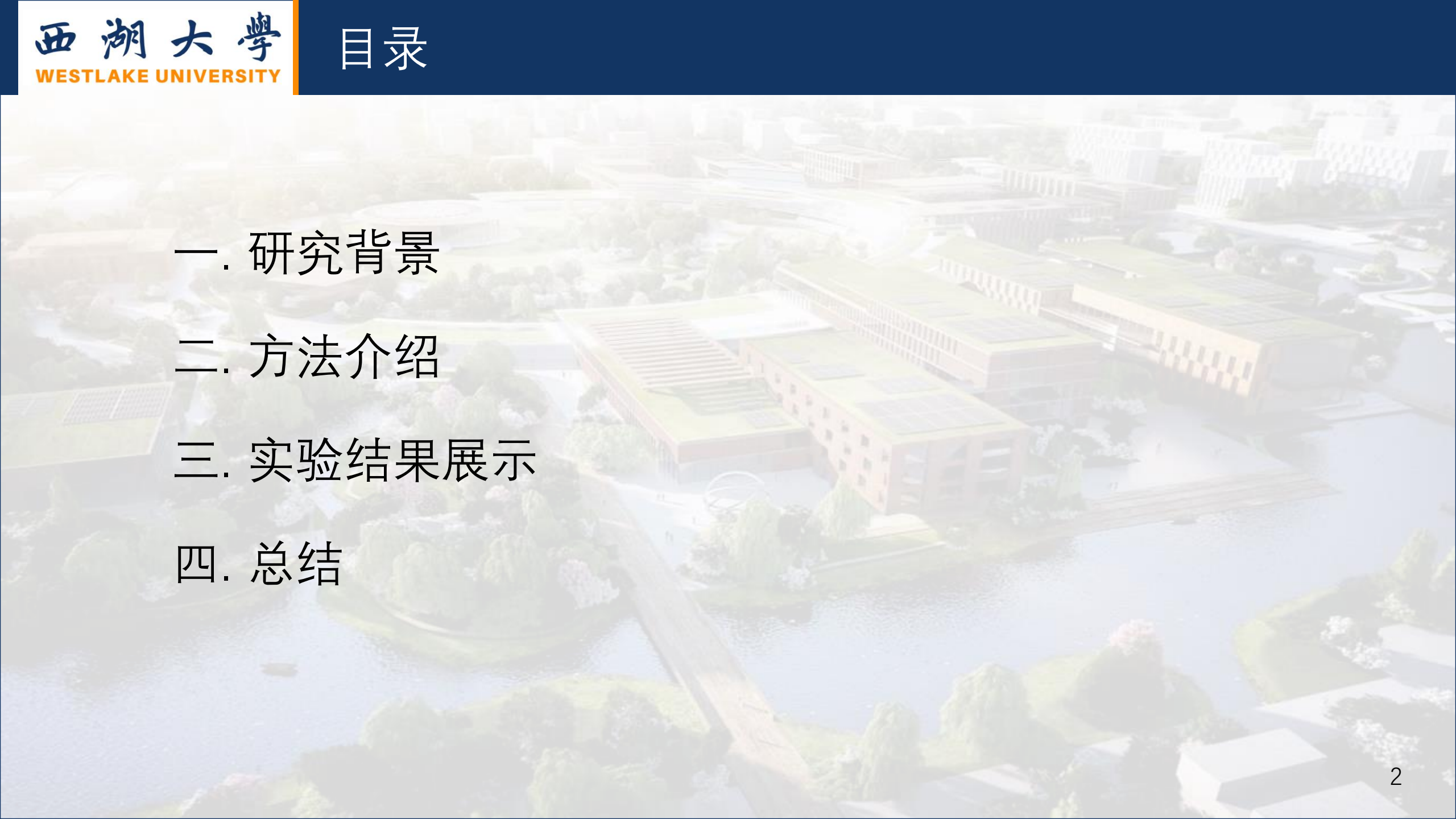
# Attributes-Guided and Pure-Visual Attention Alignment for Few-Shot Recognition

黄思腾

机器智能实验室 (MiLAB)

西湖大学工学院

2020.12.15

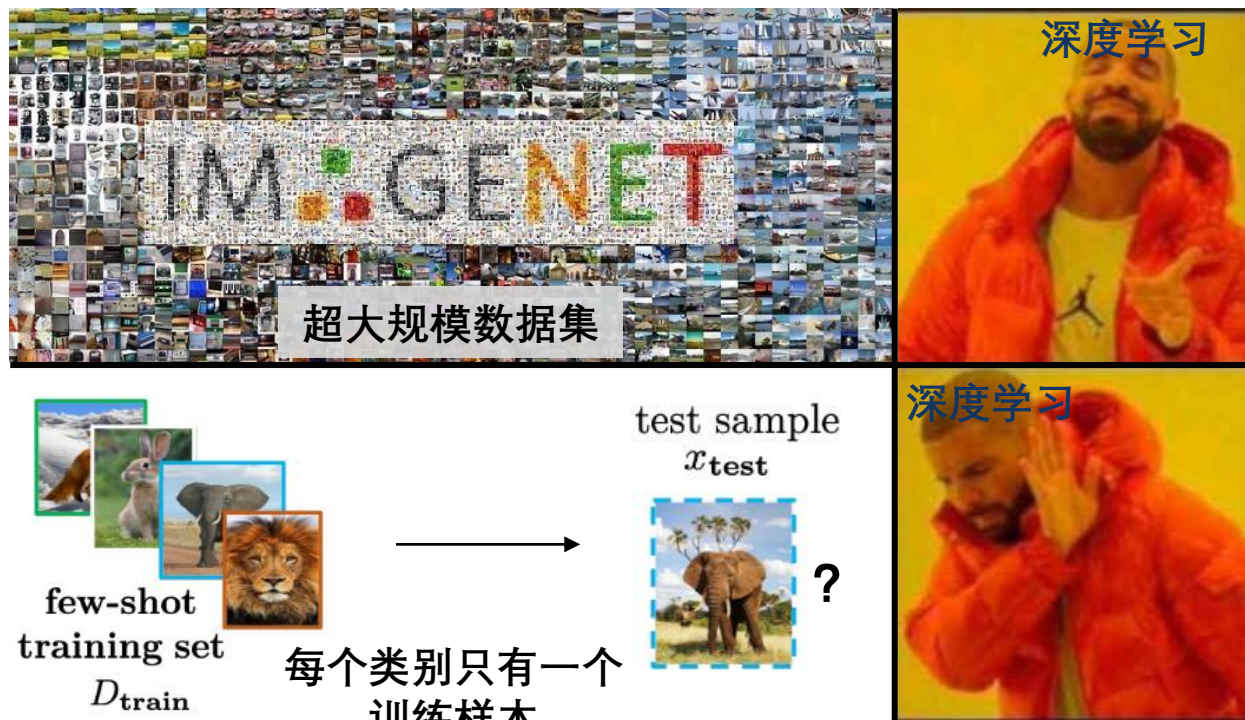
- 
- 一. 研究背景
  - 二. 方法介绍
  - 三. 实验结果展示
  - 四. 总结

一. 研究背景

二. 方法介绍

三. 实验结果展示

四. 总结







机器人学习拿起一个新的物体并将其放入元训练阶段未见过的碗中[2]

[1] Ravi, Sachin, and Hugo Larochelle. "Optimization as a model for few-shot learning." ICLR 2016.

[2] One-Shot Imitation from Watching Videos. <https://bair.berkeley.edu/blog/2018/06/28/dam/>

- 目标：完成只有少量监督信息可供使用的测试任务
- 元学习（**meta-learning**）：“学会学习”，用训练数据学习如何利用测试任务的少量支持样本来根据当前任务进行自适应（即快速学习）
- 基于度量的元学习：学习如何度量样本之间的相似度
  - MatchingNet (2016), ProtoNet (2017), TADAM (2018), DeepEMD (2020)
- 基于模型的元学习：学习生成用于快速学习的模型的参数
  - MANN (2016), MetaNet (2017), LGM-Net (2019)
- 基于优化的元学习：学习调整优化算法来快速优化模型参数
  - LSTM Meta-learner (2017), MAML (2017), Reptile (2018), LEO (2019)



教孩子学会认识一种新的物体通常伴随着语言描述等语义信息[1]。



自然语言描述

African equines with distinctive black-and-white striped coats.

类标签词嵌入

'zebra'

0.6	0.2	-0.4	...	0.1
-----	-----	------	-----	-----

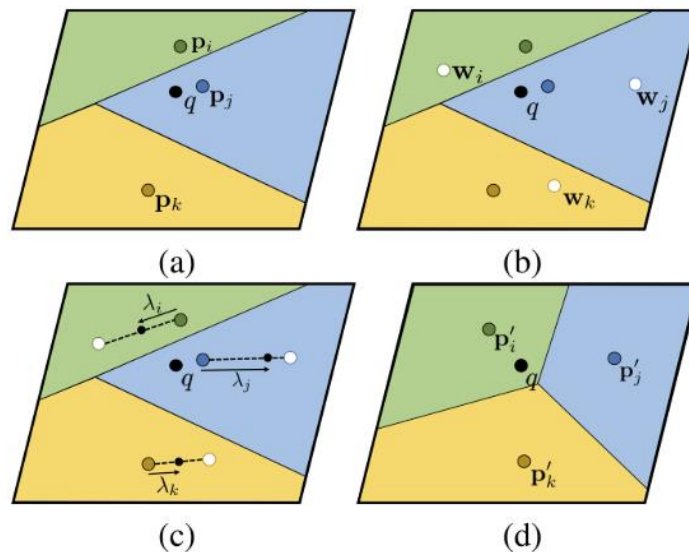
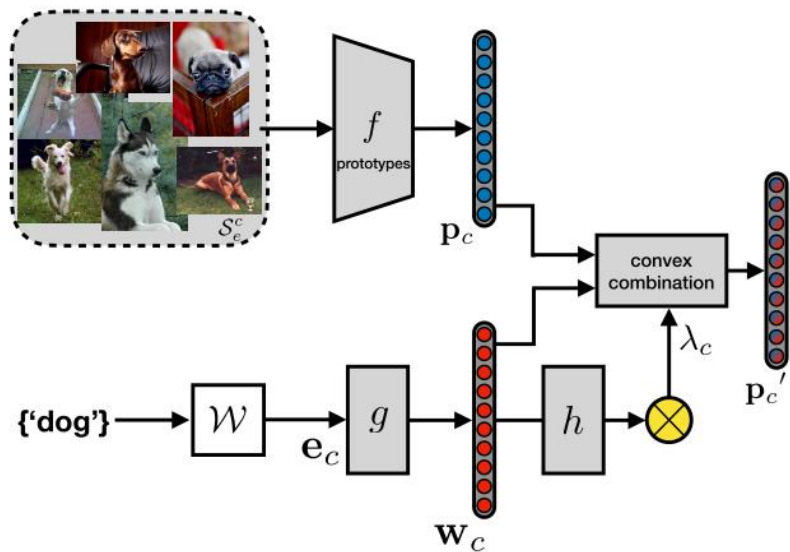
word embedding

人工属性标注

{  
'horse-like',  
'white and black stripes',  
'Mohawk-like mane',  
...  
}

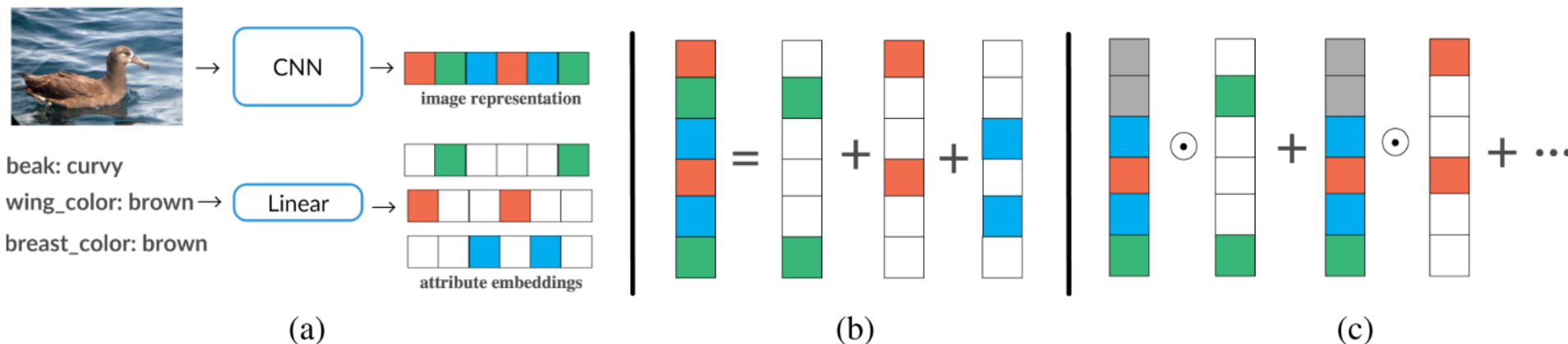
语义信息指能够更加抽象地表示图像内容的信息。





[1]将视觉类别原型与语义模态向量自适应计算比例后凸组合，作为新的类别原型。

[2]用语义模态向量对视觉特征的学习进行组合性上的约束。



[1] Xing C, Rostamzadeh N, Oreshkin B, et al. "Adaptive Cross-Modal Few-Shot Learning." NeurIPS 2019.

[2] Tokmakov, Pavel, Yu-Xiong Wang, and Martial Hebert. "Learning compositional representations for few-shot recognition." ICCV 2019.



一. 研究背景

二. 方法介绍

三. 实验结果展示

四. 总结

在只有支持样本的附加语义模态可得的设定下，现有方法

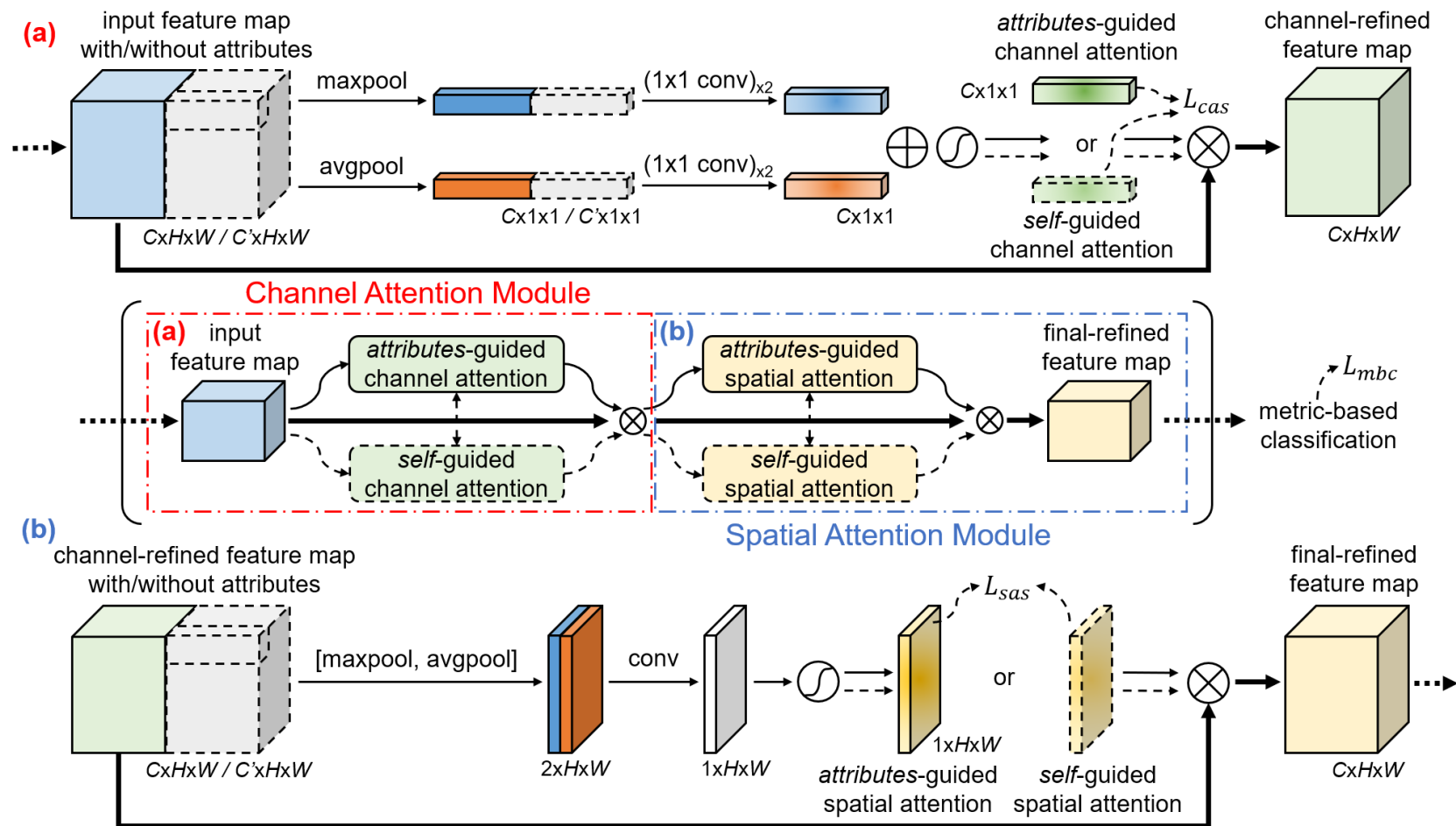
(1) 只关注约束或增强支持样本表征，而忽视了为查询样本显式设计特殊机制来优化表征；

(2) 忽视视觉和语义的特征空间天然异构，导致融合得到的支持样本表征和纯视觉的同类查询表征可能存在偏移。



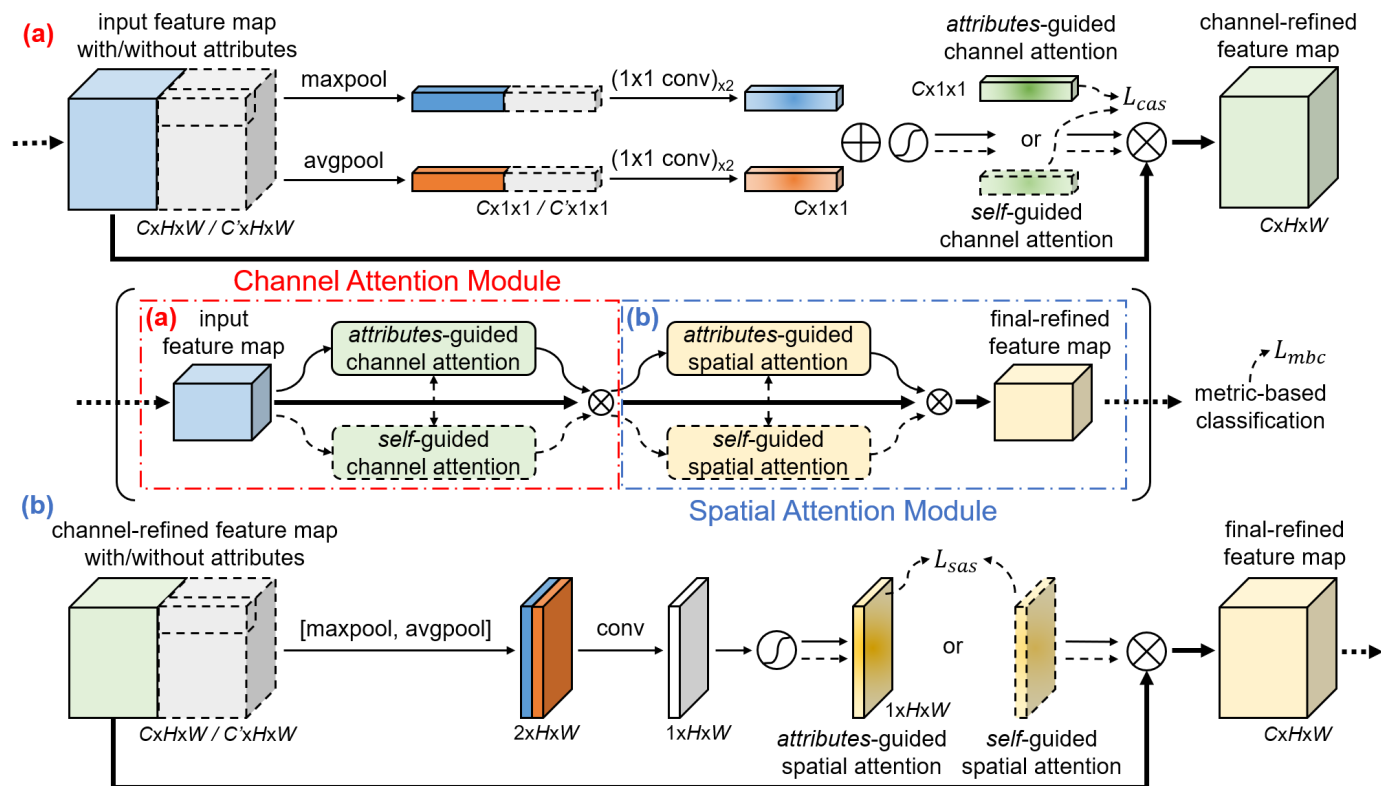
Attributes guide you to learn!

# 属性指导的注意力模块 (AGAM)



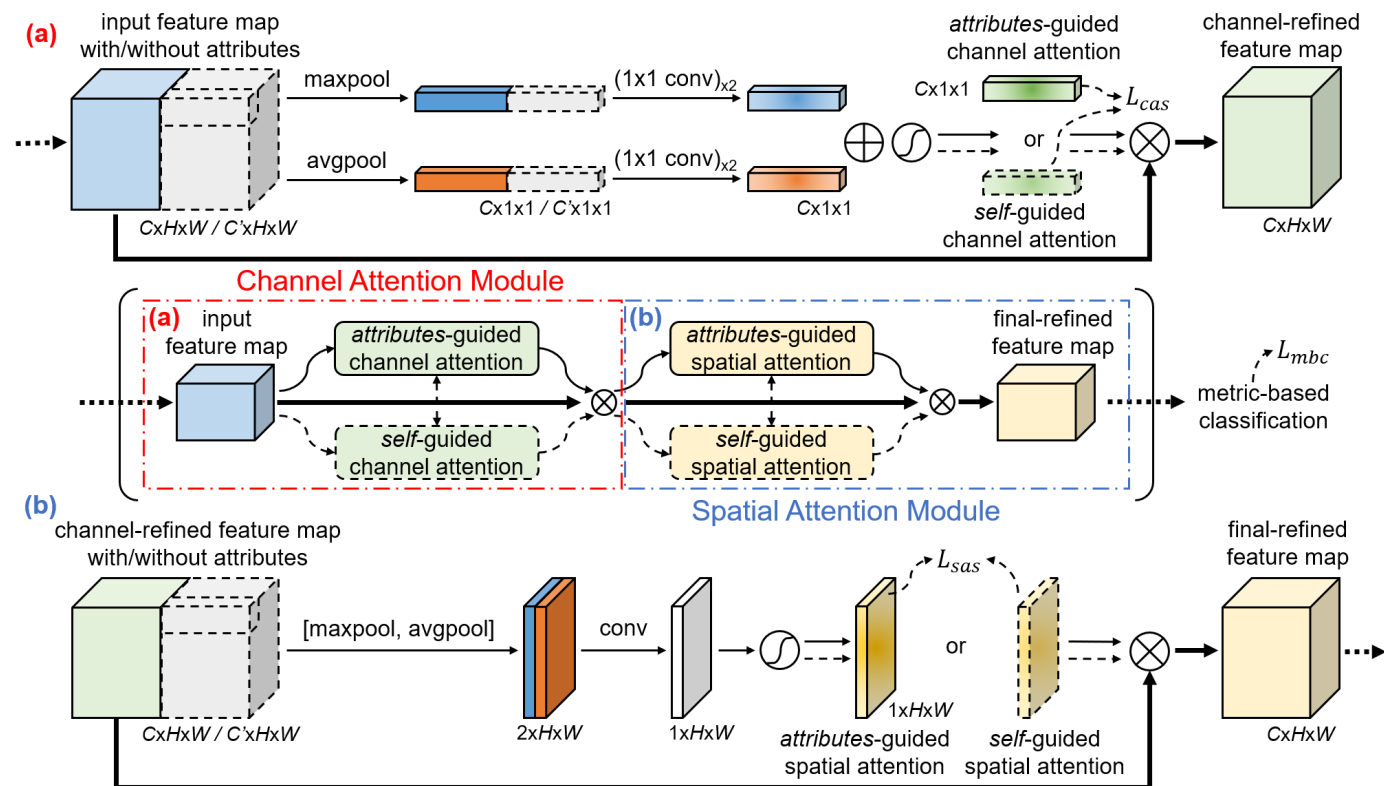
AGAM的总体架构。取决于图像的属性标注是否可得，属性指导分支和自我指导分支中的一支被选择。输入的特征依次经过 (a) 通道注意力模块和 (b) 空间注意力模块来获得最终改善的特征。

- 问题 1: 现有方法只关注约束或增强支持样本表征, 而忽视了为查询样本设计特殊机制来优化表征;
- 方案: 我们为查询样本同样设计了自我指导分支, 能够细粒度地增强查询分支的表征。

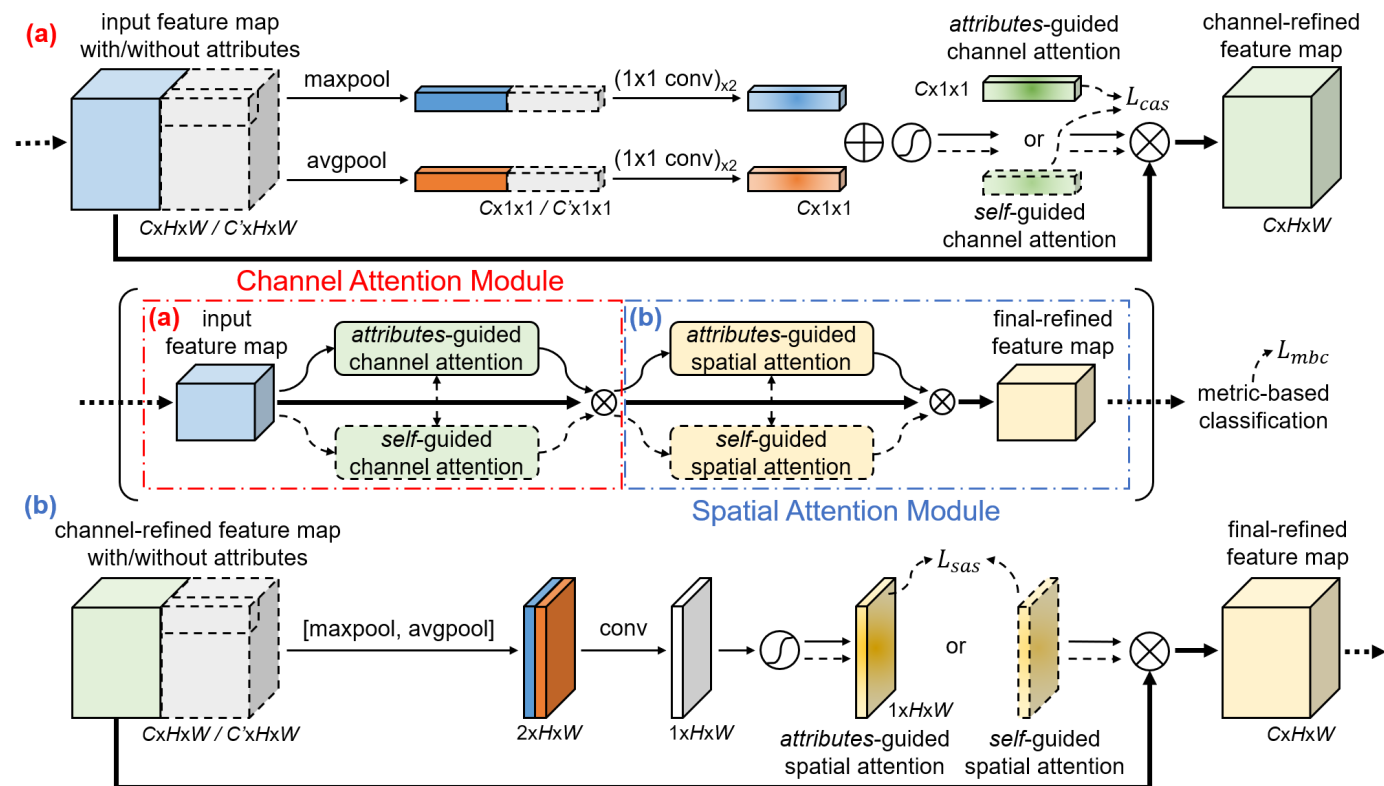




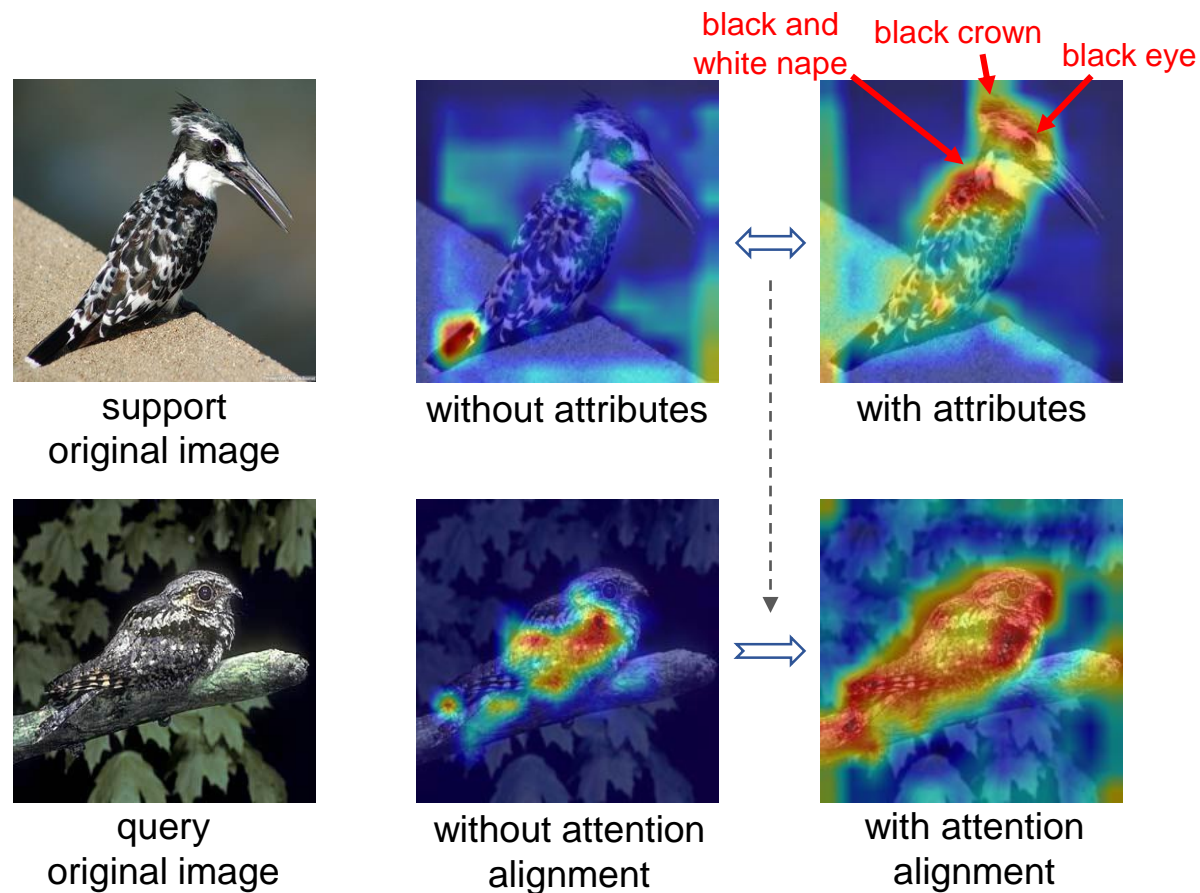
- 问题 2: 现有方法忽视视觉和语义的**特征空间天然异构**, 导致融合得到的支持样本表征和纯视觉的同类查询样本表征可能存在偏移;
- 方案: 两条分支采用相似的注意力机制在视觉特征空间中选择性加强或减弱。



- 问题 3: 对于同类样本, 是否有属性指导特征选择可能导致两条分支关注不同的通道或区域;
- 方案: 设计**注意力对齐损失**作为注意力对齐机制, 约束支持样本在经过两条分支时注意力模块关注的特征尽可能相同。



- 问题 3: 对于同类样本, 是否有属性指导特征选择可能导致两条分支关注不同的通道或区域;
- 方案: 设计**注意力对齐损失**作为注意力对齐机制, 约束支持样本在经过两条分支时注意力模块关注的特征尽可能相同。



我们提出的注意力对齐机制能够鼓励在没有属性标注时也学会关注更具语义和判别性的视觉特征。

一. 研究背景

二. 方法介绍

三. 实验结果展示

四. 总结



Method	CUB		SUN	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MatchingNet (Vinyals et al. 2016), <i>paper</i>	61.16 $\pm$ 0.89	72.86 $\pm$ 0.70	-	-
MatchingNet (Vinyals et al. 2016), <i>our implementation</i>	62.82 $\pm$ 0.36	73.22 $\pm$ 0.23	55.72 $\pm$ 0.40	76.59 $\pm$ 0.21
MatchingNet (Vinyals et al. 2016) <b>with AGAM</b>	<b>71.58 <math>\pm</math> 0.30</b> <i>+8.76</i>	<b>75.46 <math>\pm</math> 0.28</b> <i>+2.24</i>	<b>64.95 <math>\pm</math> 0.35</b> <i>+9.23</i>	<b>79.06 <math>\pm</math> 0.19</b> <i>+2.47</i>
ProtoNet (Snell, Swersky, and Zemel 2017), <i>paper</i>	51.31 $\pm$ 0.91	70.77 $\pm$ 0.69	-	-
ProtoNet (Snell, Swersky, and Zemel 2017), <i>our implementation</i>	53.01 $\pm$ 0.34	71.91 $\pm$ 0.22	57.76 $\pm$ 0.29	79.27 $\pm$ 0.19
ProtoNet (Snell, Swersky, and Zemel 2017) <b>with AGAM</b>	<b>75.87 <math>\pm</math> 0.29</b> <i>+22.86</i>	<b>81.66 <math>\pm</math> 0.25</b> <i>+9.75</i>	<b>65.15 <math>\pm</math> 0.31</b> <i>+7.39</i>	<b>80.08 <math>\pm</math> 0.21</b> <i>+0.81</i>
RelationNet (Sung et al. 2018), <i>paper</i>	62.45 $\pm$ 0.98	76.11 $\pm$ 0.69	-	-
RelationNet (Sung et al. 2018), <i>our implementation</i>	58.62 $\pm$ 0.37	78.98 $\pm$ 0.24	49.58 $\pm$ 0.35	76.21 $\pm$ 0.19
RelationNet (Sung et al. 2018) <b>with AGAM</b>	<b>66.98 <math>\pm</math> 0.31</b> <i>+8.36</i>	<b>80.33 <math>\pm</math> 0.40</b> <i>+1.35</i>	<b>59.05 <math>\pm</math> 0.32</b> <i>+9.47</i>	<b>77.52 <math>\pm</math> 0.18</b> <i>+1.31</i>

Table 1: Average accuracy (%) comparison with 95% confidence intervals before and after incorporating AGAM into existing methods using a Conv-4 backbone. Best results are displayed in **boldface**, and improvements are displayed in *italics*.

Method	Backbone	Test Accuracy	
		5-way 1-shot	5-way 5-shot
MatchingNet (Vinyals et al. 2016) <sup>†</sup>	Conv-4	55.72 $\pm$ 0.40	76.59 $\pm$ 0.21
ProtoNet (Snell, Swersky, and Zemel 2017) <sup>†</sup>	Conv-4	57.76 $\pm$ 0.29	79.27 $\pm$ 0.19
RelationNet (Sung et al. 2018) <sup>†</sup>	Conv-4	49.58 $\pm$ 0.35	76.21 $\pm$ 0.19
Comp. (Tokmakov, Wang, and Hebert 2019) *	ResNet-10	45.9	67.1
AM3 (Xing et al. 2019) <sup>†</sup> *	Conv-4	62.79 $\pm$ 0.32	79.69 $\pm$ 0.23
<b>AGAM (OURS) *</b>	Conv-4	<b>65.15 <math>\pm</math> 0.31</b>	<b>80.08 <math>\pm</math> 0.21</b>

Table 3: Average accuracy (%) comparison to state-of-the-arts with 95% confidence intervals on the SUN dataset. <sup>†</sup> denotes that it is our implementation. \* denotes that it uses auxiliary attributes. Best results are displayed in **boldface**.

Method	Backbone	Test Accuracy	
		5-way 1-shot	5-way 5-shot
MatchingNet (Vinyals et al. 2016)	Conv-4	61.16 $\pm$ 0.89	72.86 $\pm$ 0.70
ProtoNet (Snell, Swersky, and Zemel 2017)	Conv-4	51.31 $\pm$ 0.91	70.77 $\pm$ 0.69
RelationNet (Sung et al. 2018)	Conv-4	62.45 $\pm$ 0.98	76.11 $\pm$ 0.69
MACO (Hilliard et al. 2018)	Conv-4	60.76	74.96
MAML (Finn, Abbeel, and Levine 2017)	Conv-4	55.92 $\pm$ 0.95	72.09 $\pm$ 0.76
Baseline (Chen et al. 2019a)	Conv-4	47.12 $\pm$ 0.74	64.16 $\pm$ 0.71
Baseline++ (Chen et al. 2019a)	Conv-4	60.53 $\pm$ 0.83	79.34 $\pm$ 0.61
Comp. (Tokmakov, Wang, and Hebert 2019) *	ResNet-10	53.6	74.6
AM3 (Xing et al. 2019) <sup>†</sup> *	Conv-4	73.78 $\pm$ 0.28	81.39 $\pm$ 0.26
<b>AGAM (OURS) *</b>	Conv-4	<b>75.87 <math>\pm</math> 0.29</b>	<b>81.66 <math>\pm</math> 0.25</b>
MatchingNet (Vinyals et al. 2016) <sup>†</sup>	ResNet-12	60.96 $\pm$ 0.35	77.31 $\pm$ 0.25
ProtoNet (Snell, Swersky, and Zemel 2017)	ResNet-12	68.8	76.4
RelationNet (Sung et al. 2018) <sup>†</sup>	ResNet-12	60.21 $\pm$ 0.35	80.18 $\pm$ 0.25
TADAM (Oreshkin, López, and Lacoste 2018)	ResNet-12	69.2	78.6
FEAT (Ye et al. 2020)	ResNet-12	68.87 $\pm$ 0.22	82.90 $\pm$ 0.15
MAML (Finn, Abbeel, and Levine 2017)	ResNet-18	69.96 $\pm$ 1.01	82.70 $\pm$ 0.65
Baseline (Chen et al. 2019a)	ResNet-18	65.51 $\pm$ 0.87	82.85 $\pm$ 0.55
Baseline++ (Chen et al. 2019a)	ResNet-18	67.02 $\pm$ 0.90	83.58 $\pm$ 0.54
Delta-encoder (Bengio et al. 2018)	ResNet-18	69.8	82.6
Dist. ensemble (Dvornik, Mairal, and Schmid 2019)	ResNet-18	68.7	83.5
SimpleShot (Wang et al. 2019)	ResNet-18	70.28	86.37
AM3 (Xing et al. 2019) *	ResNet-12	73.6	79.9
Multiple-Semantics (Schwartz et al. 2019) * <sup>◦</sup> •	DenseNet-121	76.1	82.9
Dual TriNet (Chen et al. 2019b) * <sup>◦</sup>	ResNet-18	69.61 $\pm$ 0.46	84.10 $\pm$ 0.35
<b>AGAM (OURS) *</b>	ResNet-12	<b>79.58 <math>\pm</math> 0.25</b>	<b>87.17 <math>\pm</math> 0.23</b>

Table 2: Average accuracy (%) comparison to state-of-the-arts with 95% confidence intervals on the CUB dataset. <sup>†</sup> denotes that it is our implementation. \* denotes that it uses auxiliary attributes. <sup>◦</sup> denotes that it uses auxiliary label embeddings. • denotes that it uses auxiliary descriptions of the categories. Best results are displayed in **boldface**.

Loss Type	CUB	
	5-way 1-shot	5-way 5-shot
L1	66.95 $\pm$ 0.30	78.40 $\pm$ 0.25
MSE	69.83 $\pm$ 0.30	77.35 $\pm$ 0.22
smoothL1	72.42 $\pm$ 0.30	75.72 $\pm$ 0.31
soft margin	<b>75.87 <math>\pm</math> 0.29</b>	<b>81.66 <math>\pm</math> 0.25</b>

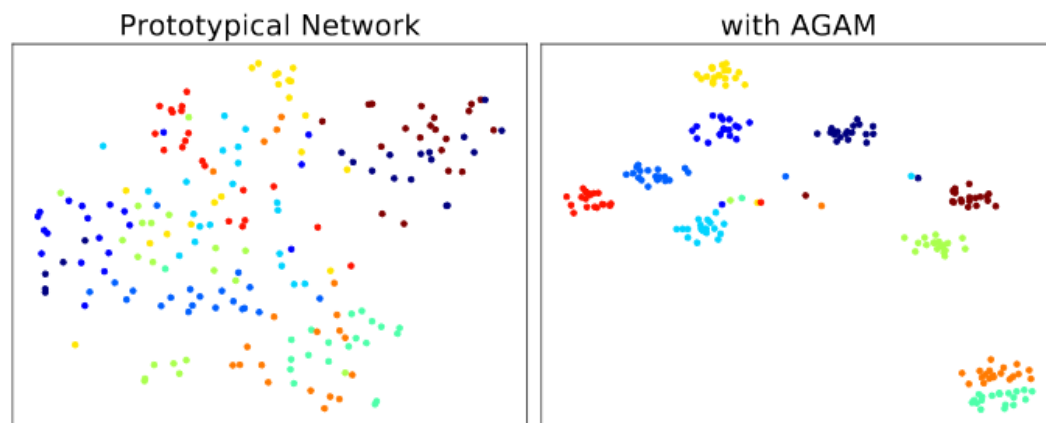
  

Loss Type	SUN	
	5-way 1-shot	5-way 5-shot
L1	60.56 $\pm$ 0.33	76.14 $\pm$ 0.26
MSE	59.54 $\pm$ 0.35	78.35 $\pm$ 0.26
smoothL1	62.07 $\pm$ 0.31	78.42 $\pm$ 0.23
soft margin	<b>65.15 <math>\pm</math> 0.31</b>	<b>80.08 <math>\pm</math> 0.21</b>

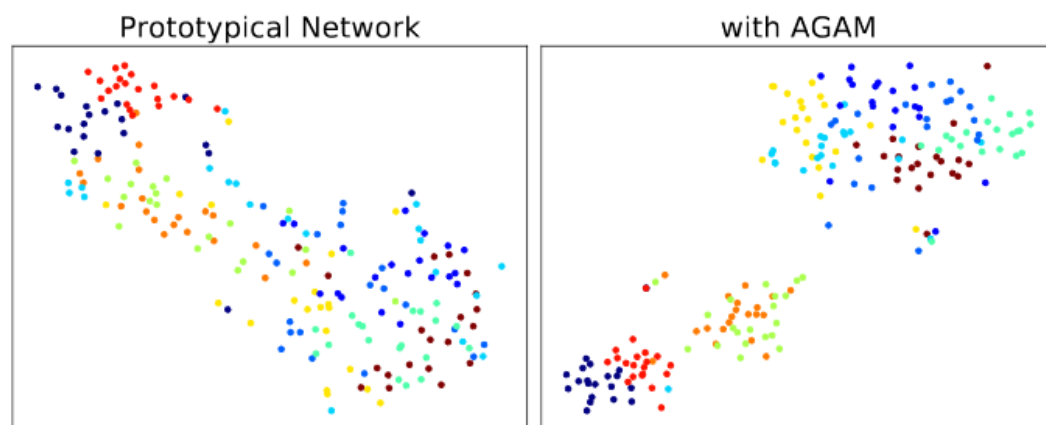
Table 1: Ablation test results of different attention alignment losses based on AGAM with a Conv-4 backbone. Average accuracies (%) with 95% confidence intervals of each model are reported. Best results are displayed in **boldface**.

Method	Test Accuracy	
	5-way 1-shot	5-way 5-shot
AGAM	<b>75.87 <math>\pm</math> 0.29</b>	<b>81.66 <math>\pm</math> 0.25</b>
AGAM_SACA	74.22 $\pm$ 0.27	79.72 $\pm$ 0.26
w/o avgpool	66.27 $\pm$ 0.29	76.58 $\pm$ 0.25
w/o maxpool	67.60 $\pm$ 0.29	77.09 $\pm$ 0.22
w/o CA	54.91 $\pm$ 0.36	80.52 $\pm$ 0.24
w/o SA	69.66 $\pm$ 0.31	76.24 $\pm$ 0.27
w/o $L_{cas}$	74.88 $\pm$ 0.26	77.78 $\pm$ 0.26
w/o $L_{sas}$	74.29 $\pm$ 0.27	77.87 $\pm$ 0.23
w/o $L_{cas}$ & $L_{sas}$	75.37 $\pm$ 0.31	78.92 $\pm$ 0.27

Table 3: Ablation test results of AGAM on CUB. Average accuracies (%) with 95% confidence intervals of each model are reported. Best results are displayed in **boldface**.



(a) Results on the CUB dataset.

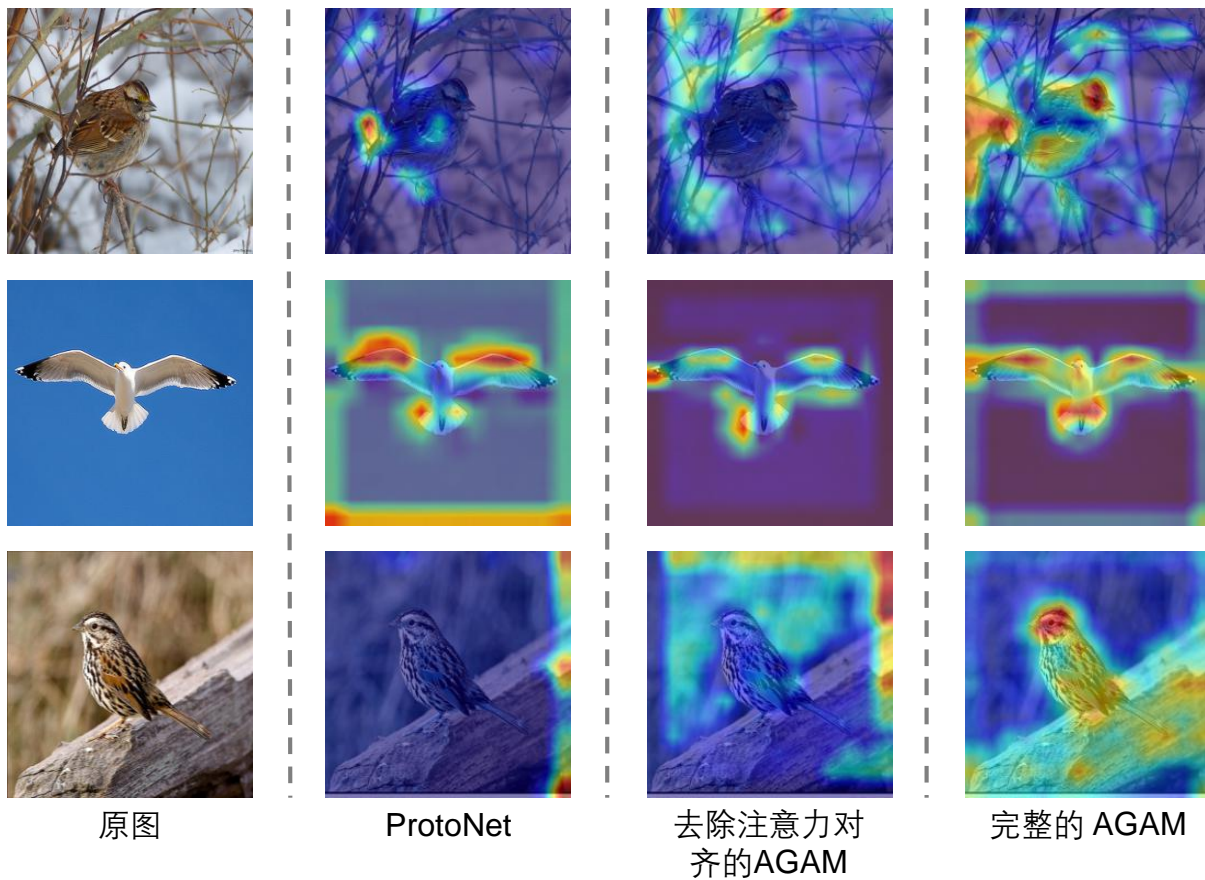


(b) Results on the SUN dataset.

原型网络（Prototypical Network）在加上我们提出的 AGAM 前后特征表征的 t-SNE 可视化对比。


加上 AGAM 会使得同类特征形成更紧密和类间可分离的簇。这表明在属性指导下学到的特征更有区分性。





**查询**样本的梯度加权类别激活映射 (Grad-CAM) 可视化。

完整的 AGAM 比去掉注意力对齐机制的 AGAM 能够注意到更具代表性的区域，说明 **自我指导分支能够受益于注意力对齐机制**。

- 
- 一. 研究背景
  - 二. 方法介绍
  - 三. 实验结果展示
  - 四. 总结**

- 用恰当的方式利用**辅助语义模态**有助于提升**小样本识别**的效果。
- 我们使用了**通道注意力**和**空间注意力**来学习应该被加强或抑制的信息。在提升视觉表征的信息量和可区分性的同时，视觉内容和对应属性协同提取的特征和纯视觉特征处于同一空间。
- 我们提出一种属性指导分支和自我指导分支间的**注意力对齐机制**，使得属性指导分支的监督信号鼓励自我指导分支在没有属性时也能够关注更重要的特征。
- 我们设计了大量的实验，来证明我们的**轻量**模块能够大幅度提升各种基于度量的小样本学习方法的性能，并在多个数据集上达到**最好效果**。

# 谢谢大家！

## Q&A

