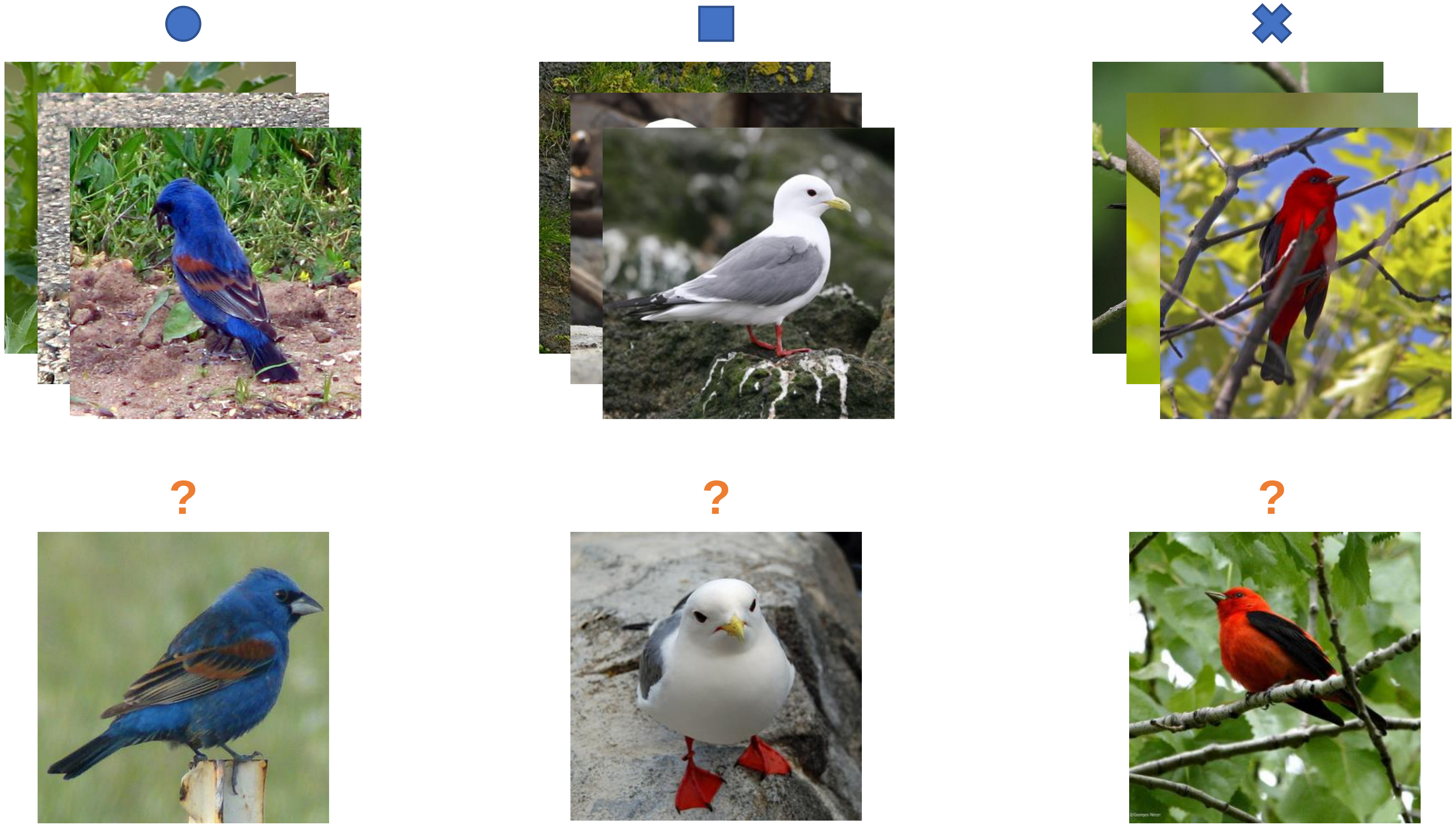
Attributes-Guided and Pure-Visual Attention Alignment for Few-Shot Recognition

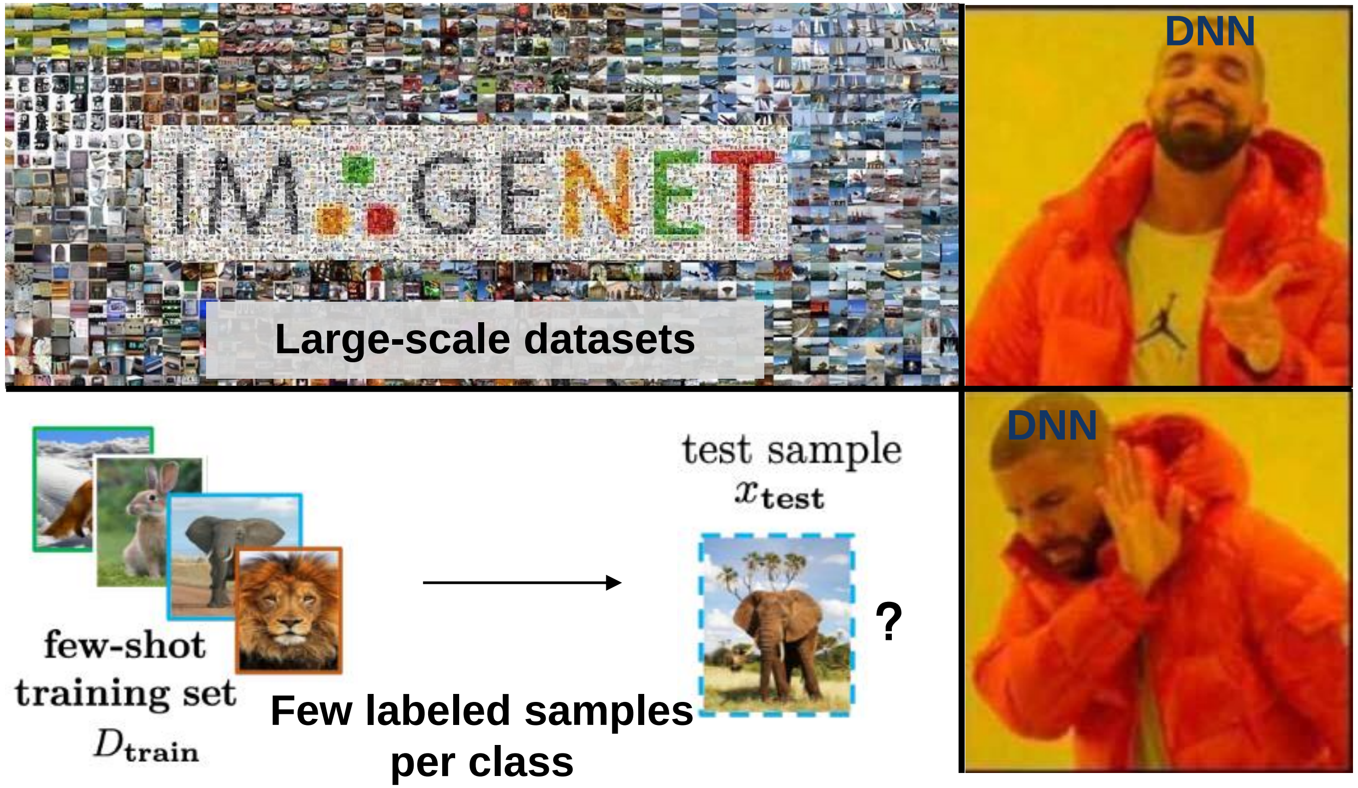
Siteng Huang, Min Zhang, Yachen Kang and Donglin Wang*

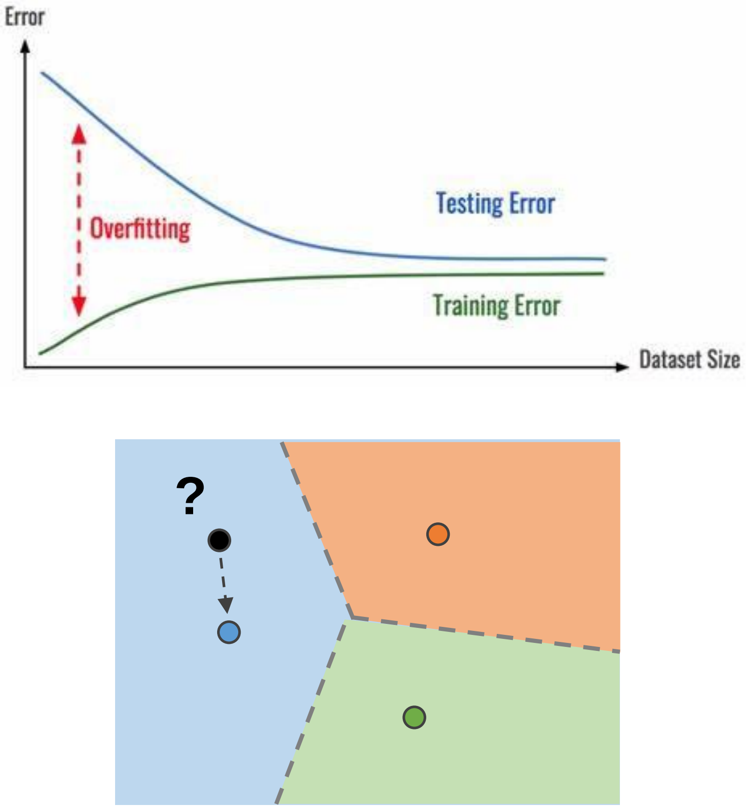
WESTLAKE UNIVERSITY

Challenge of few-shot recognition



Challenge of few-shot recognition





Learning with auxiliary semantic modalities

language description → African equines with distinctive black-and-white striped coats.

label embedding → 'zebra' →

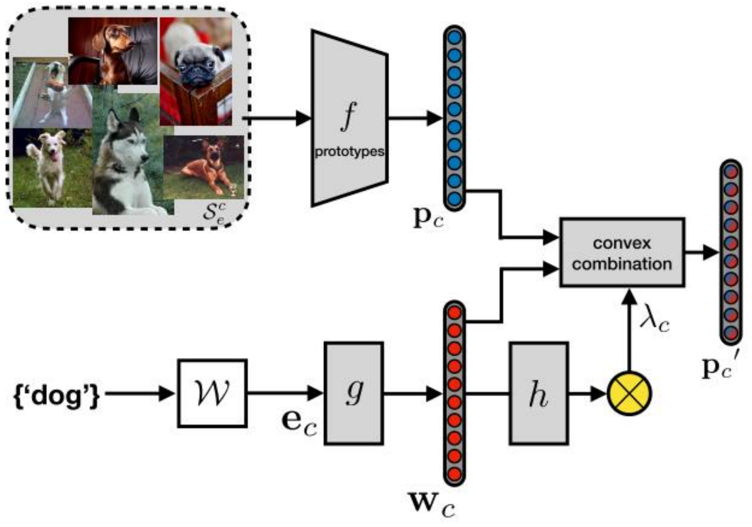
0.6	0.2	-0.4	…	0.1

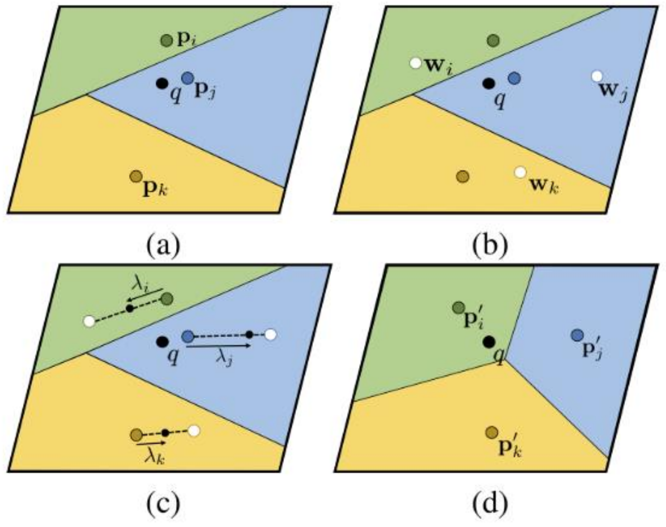
word embedding

human-annotated attributes → { 'horse-like', 'white and black stripes', 'Mohawk-like mane', … }

Semantic modalities refer to the modalities that can more abstractly represent image content.

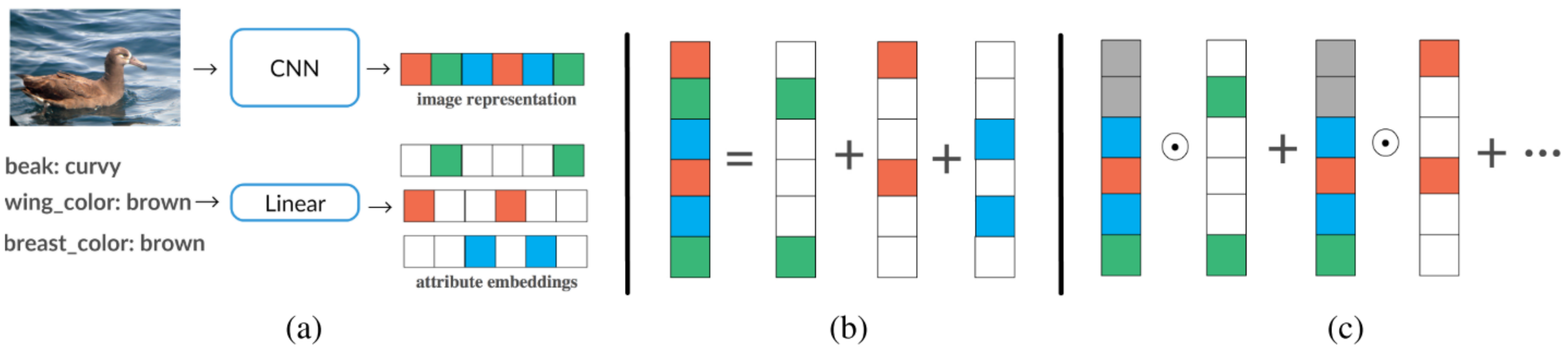
Learning with auxiliary semantic modalities





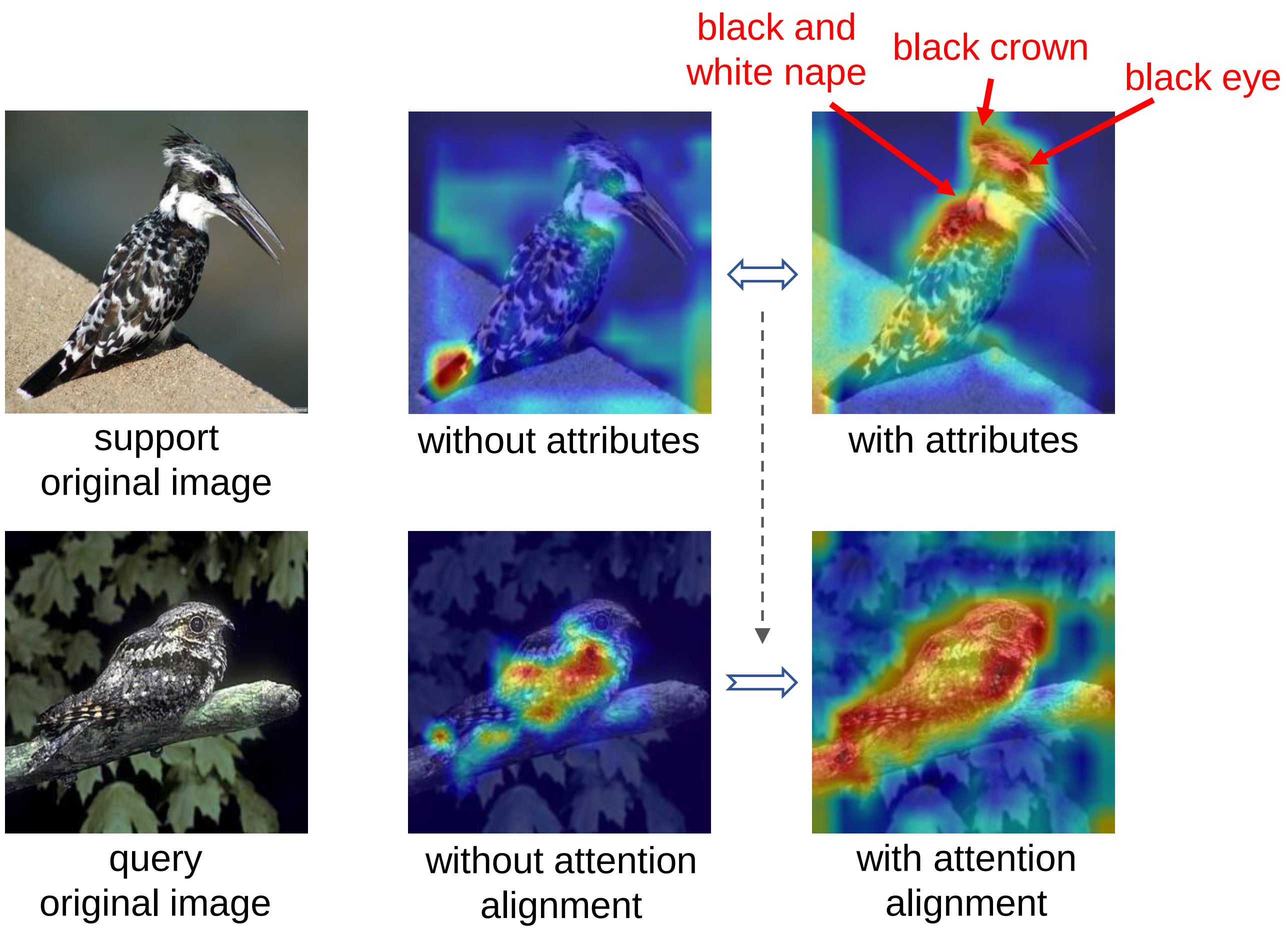
[1] models the class representation (prototype) as a convex combination of the two modalities.

[2] proposes an attribute-based regularization approach to learn compositional image representations.



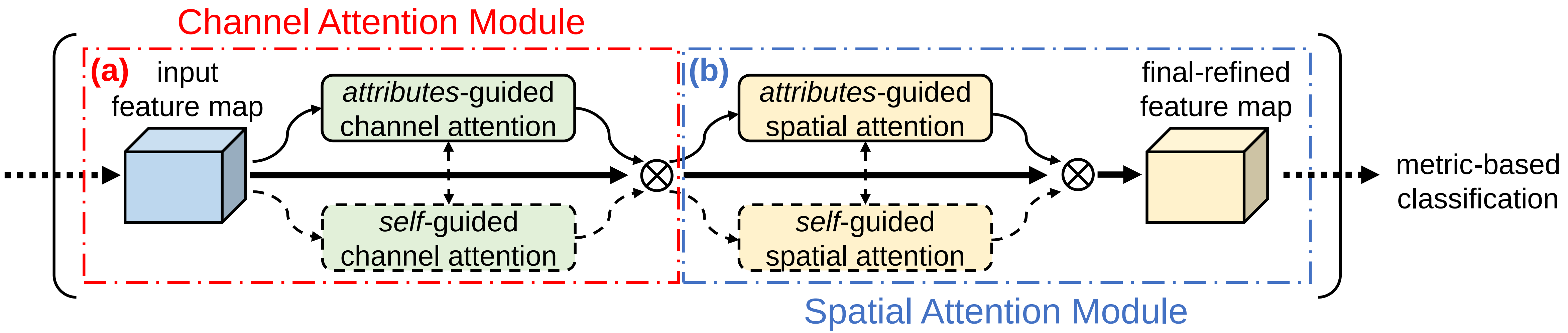
[1] Xing C, Rostamzadeh N, Oreshkin B, et al. “Adaptive cross-modal few-shot learning.” NeurIPS 2019.
[2] Tokmakov, Pavel, Yu-Xiong Wang, and Martial Hebert. "Learning compositional representations for few-shot recognition." ICCV 2019.

Attributes-guided attention module (AGAM)



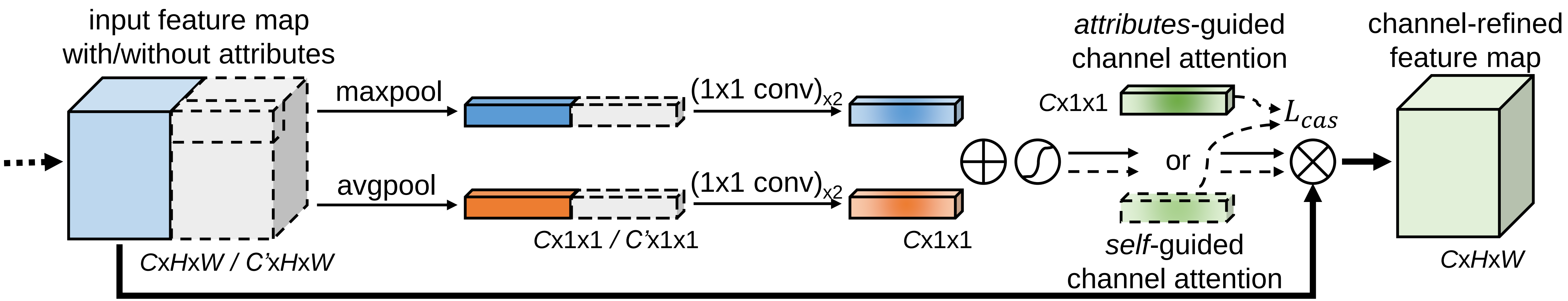
The network learns to focus on more discriminative features of both support and query samples with only attributes of support samples.

AGAM oveview



The overall framework of AGAM. Based on whether attributes to the image are available (i.e., support or query), one of the two branches is selected.

Channel Attention Module



Initial feature map: $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$

Attributes vector: $\mathbf{a} \in \mathbb{R}^{D}$

Attributes tensor: $\mathbf{A} \in \mathbb{R}^{D \times H \times W}$

The input of *attributes*-guided branch:

$$\mathbf{F}^{ag}_{c\_inp} = [\mathbf{F}; \mathbf{A}] \in \mathbb{R}^{C' \times H \times W},$$

The input of *self*-guided branch:

$$\mathbf{F}^{sg}_{c\_inp} = \mathbf{F},$$

Attributes-guided channel attention:

$$\begin{aligned}\mathbf{M}^{ag}_{c} =& \sigma(\mathbf{W}^{ag}_{1}(\mathbf{W}^{ag}_{0}(\mathrm{MaxPool}(\mathbf{F}^{ag}_{c\_inp}))) \\ &+ \mathbf{W}^{ag}_{1}(\mathbf{W}^{ag}_{0}(\mathrm{AvgPool}(\mathbf{F}^{ag}_{c\_inp})))),\end{aligned}$$

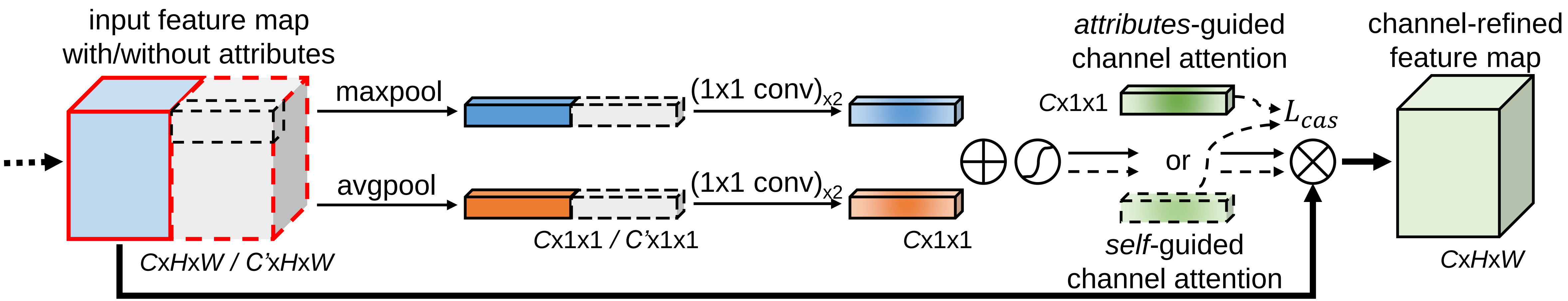
$$\mathbf{F}^{ag}_{c\_out} = \mathbf{M}^{ag}_{c} \otimes \mathbf{F},$$

Self-guided channel attention:

$$\begin{aligned}\mathbf{M}^{sg}_{c} =& \sigma(\mathbf{W}^{sg}_{1}(\mathbf{W}^{sg}_{0}(\mathrm{MaxPool}(\mathbf{F}^{sg}_{c\_inp}))) \\ &+ \mathbf{W}^{sg}_{1}(\mathbf{W}^{sg}_{0}(\mathrm{AvgPool}(\mathbf{F}^{sg}_{c\_inp})))),\end{aligned}$$

$$\mathbf{F}^{sg}_{c\_out} = \mathbf{M}^{sg}_{c} \otimes \mathbf{F}.$$

Channel Attention Module



Initial feature map: $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$

Attributes vector: $\mathbf{a} \in \mathbb{R}^{D}$

Attributes tensor: $\mathbf{A} \in \mathbb{R}^{D \times H \times W}$

The input of *attributes*-guided branch:

$$\mathbf{F}^{ag}_{c\_inp} = [\mathbf{F}; \mathbf{A}] \in \mathbb{R}^{C' \times H \times W},$$

The input of *self*-guided branch:

$$\mathbf{F}^{sg}_{c\_inp} = \mathbf{F},$$

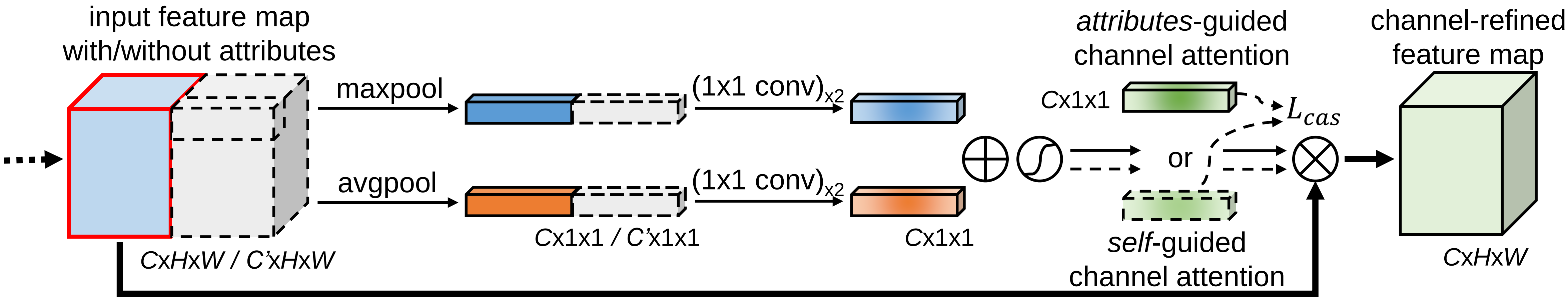
Attributes-guided channel attention:

$$\begin{aligned}\mathbf{M}^{ag}_{c} =& \sigma(\mathbf{W}^{ag}_{1}(\mathbf{W}^{ag}_{0}(\mathrm{MaxPool}(\mathbf{F}^{ag}_{c\_inp}))) \\ &+ \mathbf{W}^{ag}_{1}(\mathbf{W}^{ag}_{0}(\mathrm{AvgPool}(\mathbf{F}^{ag}_{c\_inp})))), \\ \mathbf{F}^{ag}_{c\_out} =& \mathbf{M}^{ag}_{c} \otimes \mathbf{F},\end{aligned}$$

Self-guided channel attention:

$$\begin{aligned}\mathbf{M}^{sg}_{c} =& \sigma(\mathbf{W}^{sg}_{1}(\mathbf{W}^{sg}_{0}(\mathrm{MaxPool}(\mathbf{F}^{sg}_{c\_inp}))) \\ &+ \mathbf{W}^{sg}_{1}(\mathbf{W}^{sg}_{0}(\mathrm{AvgPool}(\mathbf{F}^{sg}_{c\_inp})))), \\ \mathbf{F}^{sg}_{c\_out} =& \mathbf{M}^{sg}_{c} \otimes \mathbf{F}.\end{aligned}$$

Channel Attention Module



Initial feature map: $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$

Attributes vector: $\mathbf{a} \in \mathbb{R}^{D}$

Attributes tensor: $\mathbf{A} \in \mathbb{R}^{D \times H \times W}$

The input of *attributes*-guided branch:

$$\mathbf{F}^{ag}_{c\_inp} = [\mathbf{F}; \mathbf{A}] \in \mathbb{R}^{C' \times H \times W},$$

The input of *self*-guided branch:

$$\mathbf{F}^{sg}_{c\_inp} = \mathbf{F},$$

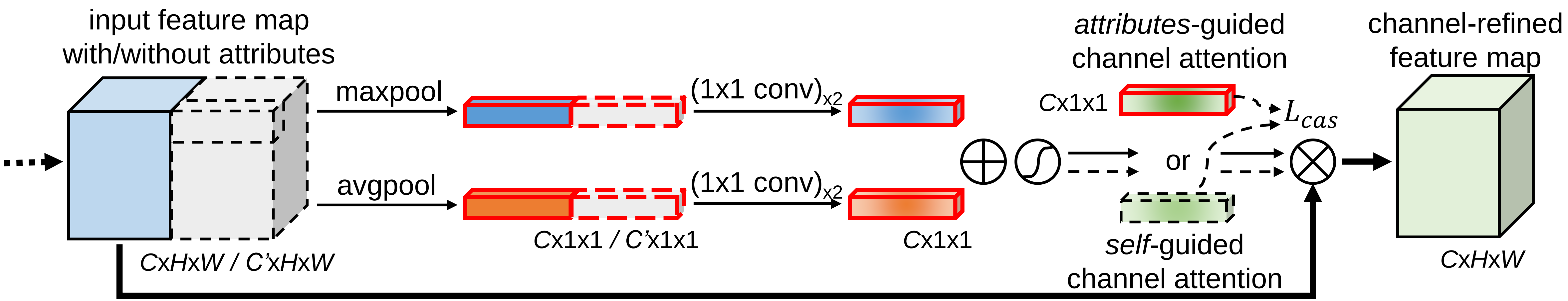
Attributes-guided channel attention:

$$\begin{aligned} \mathbf{M}^{ag}_{c} =& \sigma(\mathbf{W}^{ag}_{1}(\mathbf{W}^{ag}_{0}(\mathrm{MaxPool}(\mathbf{F}^{ag}_{c\_inp}))) \\ & + \mathbf{W}^{ag}_{1}(\mathbf{W}^{ag}_{0}(\mathrm{AvgPool}(\mathbf{F}^{ag}_{c\_inp})))), \\ \mathbf{F}^{ag}_{c\_out} =& \mathbf{M}^{ag}_{c} \otimes \mathbf{F}, \end{aligned}$$

Self-guided channel attention:

$$\begin{aligned} \mathbf{M}^{sg}_{c} =& \sigma(\mathbf{W}^{sg}_{1}(\mathbf{W}^{sg}_{0}(\mathrm{MaxPool}(\mathbf{F}^{sg}_{c\_inp}))) \\ & + \mathbf{W}^{sg}_{1}(\mathbf{W}^{sg}_{0}(\mathrm{AvgPool}(\mathbf{F}^{sg}_{c\_inp})))), \\ \mathbf{F}^{sg}_{c\_out} =& \mathbf{M}^{sg}_{c} \otimes \mathbf{F}. \end{aligned}$$

Channel Attention Module



Initial feature map: $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$

Attributes vector: $\mathbf{a} \in \mathbb{R}^{D}$

Attributes tensor: $\mathbf{A} \in \mathbb{R}^{D \times H \times W}$

The input of *attributes*-guided branch:

$$\mathbf{F}^{ag}_{c\_inp} = [\mathbf{F}; \mathbf{A}] \in \mathbb{R}^{C' \times H \times W},$$

The input of *self*-guided branch:

$$\mathbf{F}^{sg}_{c\_inp} = \mathbf{F},$$

Attributes-guided channel attention:

$$\begin{aligned}\mathbf{M}^{ag}_{c} =& \sigma(\mathbf{W}^{ag}_{1}(\mathbf{W}^{ag}_{0}(\mathrm{MaxPool}(\mathbf{F}^{ag}_{c\_inp}))) \\ &+ \mathbf{W}^{ag}_{1}(\mathbf{W}^{ag}_{0}(\mathrm{AvgPool}(\mathbf{F}^{ag}_{c\_inp})))),\end{aligned}$$

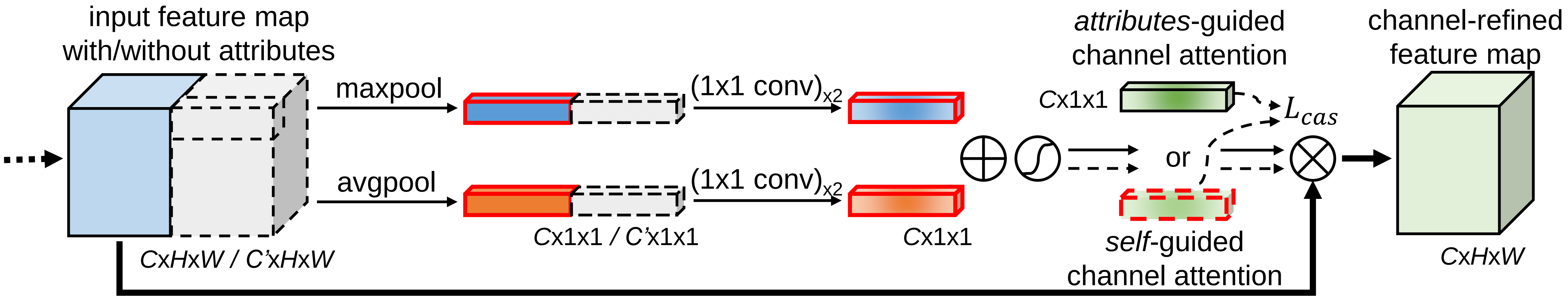
$$\mathbf{F}^{ag}_{c\_out} = \mathbf{M}^{ag}_{c} \otimes \mathbf{F},$$

Self-guided channel attention:

$$\begin{aligned}\mathbf{M}^{sg}_{c} =& \sigma(\mathbf{W}^{sg}_{1}(\mathbf{W}^{sg}_{0}(\mathrm{MaxPool}(\mathbf{F}^{sg}_{c\_inp}))) \\ &+ \mathbf{W}^{sg}_{1}(\mathbf{W}^{sg}_{0}(\mathrm{AvgPool}(\mathbf{F}^{sg}_{c\_inp})))),\end{aligned}$$

$$\mathbf{F}^{sg}_{c\_out} = \mathbf{M}^{sg}_{c} \otimes \mathbf{F}.$$

Channel Attention Module



Initial feature map: $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$

Attributes vector: $\mathbf{a} \in \mathbb{R}^{D}$

Attributes tensor: $\mathbf{A} \in \mathbb{R}^{D \times H \times W}$

The input of *attributes*-guided branch:

$$\mathbf{F}^{ag}_{c\_inp} = [\mathbf{F}; \mathbf{A}] \in \mathbb{R}^{C' \times H \times W},$$

The input of *self*-guided branch:

$$\mathbf{F}^{sg}_{c\_inp} = \mathbf{F},$$

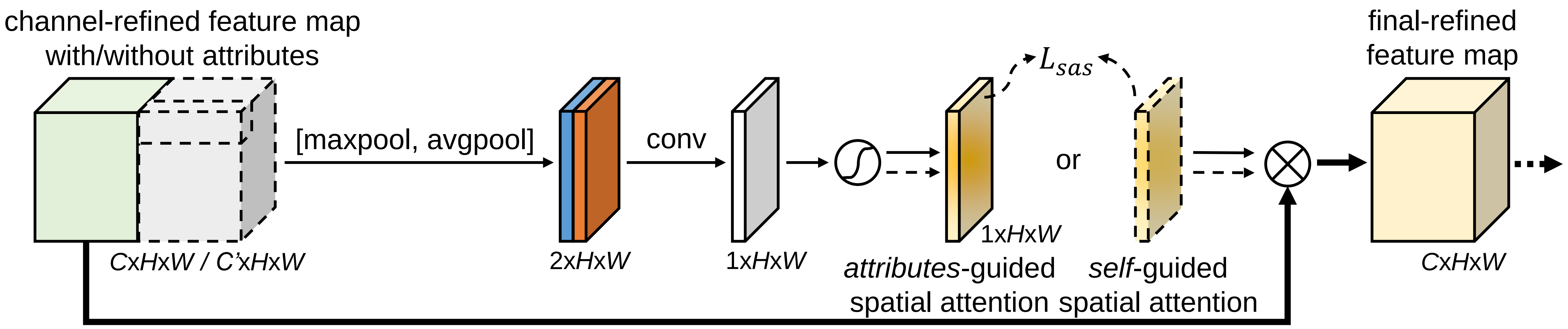
Attributes-guided channel attention:

$$\begin{aligned}\mathbf{M}^{ag}_{c} =& \sigma(\mathbf{W}^{ag}_{1}(\mathbf{W}^{ag}_{0}(\mathrm{MaxPool}(\mathbf{F}^{ag}_{c\_inp}))) \\ &+ \mathbf{W}^{ag}_{1}(\mathbf{W}^{ag}_{0}(\mathrm{AvgPool}(\mathbf{F}^{ag}_{c\_inp})))), \\ \mathbf{F}^{ag}_{c\_out} =& \mathbf{M}^{ag}_{c} \otimes \mathbf{F},\end{aligned}$$

Self-guided channel attention:

$$\begin{aligned}\mathbf{M}^{sg}_{c} =& \sigma(\mathbf{W}^{sg}_{1}(\mathbf{W}^{sg}_{0}(\mathrm{MaxPool}(\mathbf{F}^{sg}_{c\_inp}))) \\ &+ \mathbf{W}^{sg}_{1}(\mathbf{W}^{sg}_{0}(\mathrm{AvgPool}(\mathbf{F}^{sg}_{c\_inp})))), \\ \mathbf{F}^{sg}_{c\_out} =& \mathbf{M}^{sg}_{c} \otimes \mathbf{F}.\end{aligned}$$

Spatial Attention Module



The input of *attributes*-guided branch:

$$\mathbf{F}^{ag}_{s\_inp} = [\mathbf{F}^{ag}_{c\_out}; \mathbf{A}] \in \mathbb{R}^{C' \times H \times W}$$

Attributes-guided spatial attention:

$$\mathbf{M}^{ag}_{s} = \sigma(f^{ag}([\mathrm{AvgPool}(\mathbf{F}^{ag}_{s\_inp}); \mathrm{MaxPool}(\mathbf{F}^{ag}_{s\_inp})])),$$

$$\mathbf{F}^{ag}_{s\_out} = \mathbf{M}^{ag}_{s} \otimes \mathbf{F}^{ag}_{c\_out},$$

The input of *self*-guided branch:

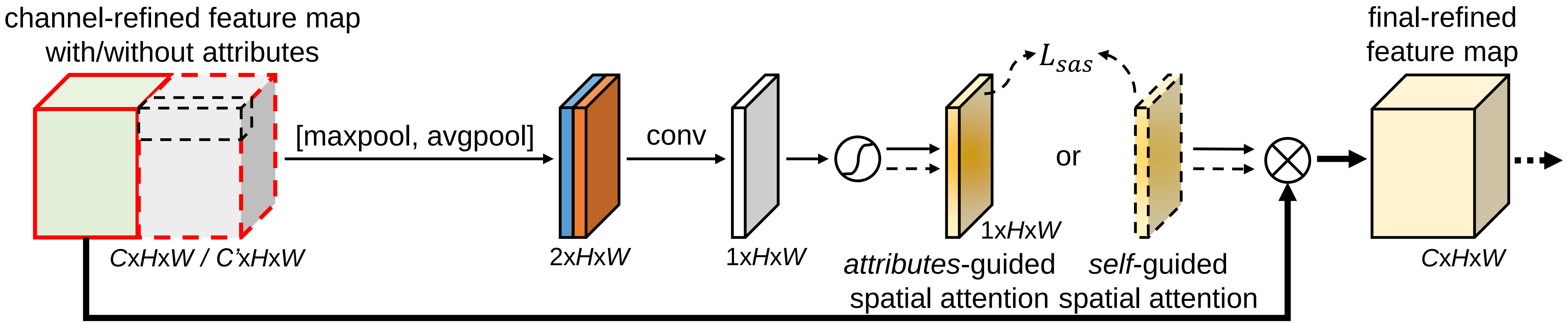
$$\mathbf{F}^{sg}_{s\_inp} = \mathbf{F}^{sg}_{c\_out} \in \mathbb{R}^{C \times H \times W}$$

Self-guided spatial attention:

$$\mathbf{M}^{sg}_{s} = \sigma(f^{sg}([\mathrm{AvgPool}(\mathbf{F}^{sg}_{s\_inp}); \mathrm{MaxPool}(\mathbf{F}^{sg}_{s\_inp})])),$$

$$\mathbf{F}^{sg}_{s\_out} = \mathbf{M}^{sg}_{s} \otimes \mathbf{F}^{sg}_{c\_out}.$$

Spatial Attention Module



The input of *attributes*-guided branch:

$$\mathbf{F}^{ag}_{s\_inp} = [\mathbf{F}^{ag}_{c\_out}; \mathbf{A}] \in \mathbb{R}^{C' \times H \times W}$$

Attributes-guided spatial attention:

$$\mathbf{M}^{ag}_{s} = \sigma(f^{ag}([\mathrm{AvgPool}(\mathbf{F}^{ag}_{s\_inp}); \mathrm{MaxPool}(\mathbf{F}^{ag}_{s\_inp})])),$$

$$\mathbf{F}^{ag}_{s\_out} = \mathbf{M}^{ag}_{s} \otimes \mathbf{F}^{ag}_{c\_out},$$

The input of *self*-guided branch:

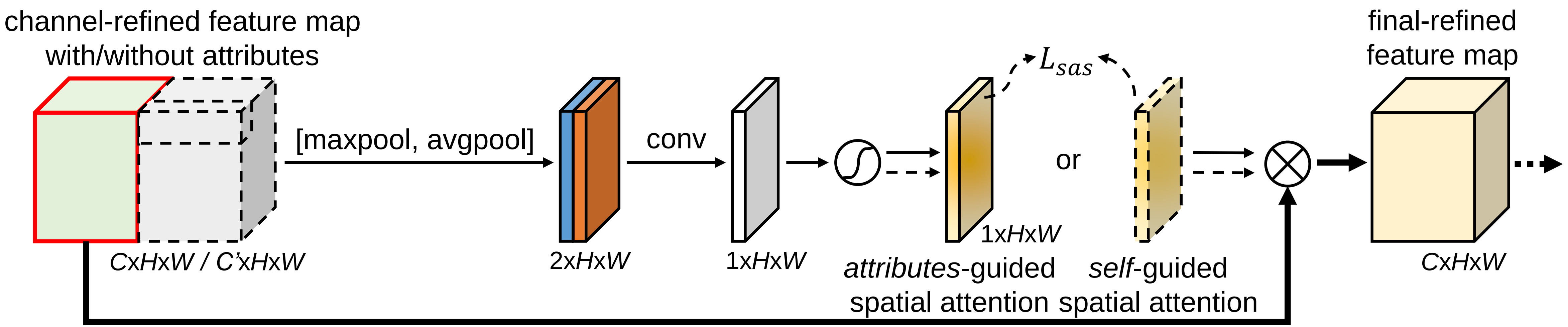
$$\mathbf{F}^{sg}_{s\_inp} = \mathbf{F}^{sg}_{c\_out} \in \mathbb{R}^{C \times H \times W}$$

Self-guided spatial attention:

$$\mathbf{M}^{sg}_{s} = \sigma(f^{sg}([\mathrm{AvgPool}(\mathbf{F}^{sg}_{s\_inp}); \mathrm{MaxPool}(\mathbf{F}^{sg}_{s\_inp})])),$$

$$\mathbf{F}^{sg}_{s\_out} = \mathbf{M}^{sg}_{s} \otimes \mathbf{F}^{sg}_{c\_out}.$$

Spatial Attention Module



The input of *attributes*-guided branch:

$$\mathbf{F}^{ag}_{s\_inp} = [\mathbf{F}^{ag}_{c\_out}; \mathbf{A}] \in \mathbb{R}^{C' \times H \times W}$$

Attributes-guided spatial attention:

$$\mathbf{M}^{ag}_{s} = \sigma(f^{ag}([\mathrm{AvgPool}(\mathbf{F}^{ag}_{s\_inp}); \mathrm{MaxPool}(\mathbf{F}^{ag}_{s\_inp})])),$$

$$\mathbf{F}^{ag}_{s\_out} = \mathbf{M}^{ag}_{s} \otimes \mathbf{F}^{ag}_{c\_out},$$

The input of *self*-guided branch:

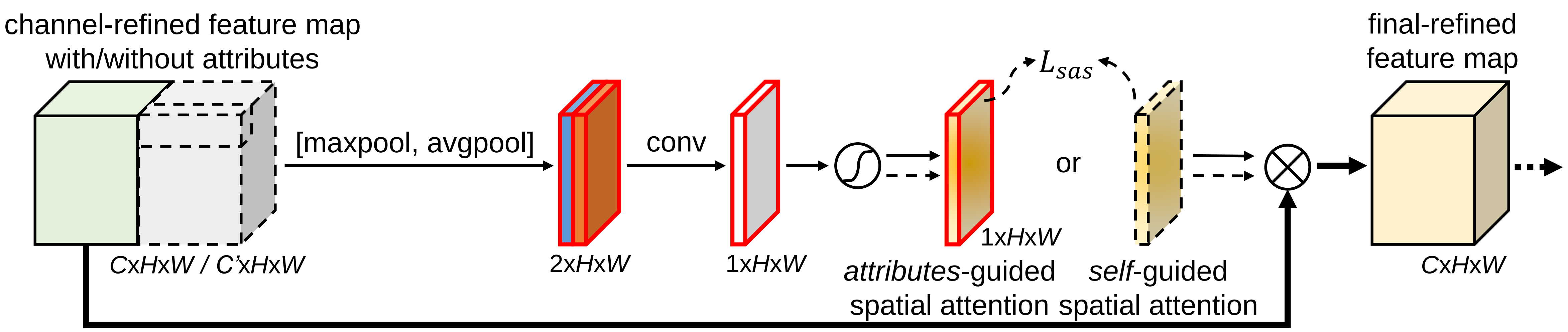
$$\mathbf{F}^{sg}_{s\_inp} = \mathbf{F}^{sg}_{c\_out} \in \mathbb{R}^{C \times H \times W}$$

Self-guided spatial attention:

$$\mathbf{M}^{sg}_{s} = \sigma(f^{sg}([\mathrm{AvgPool}(\mathbf{F}^{sg}_{s\_inp}); \mathrm{MaxPool}(\mathbf{F}^{sg}_{s\_inp})])),$$

$$\mathbf{F}^{sg}_{s\_out} = \mathbf{M}^{sg}_{s} \otimes \mathbf{F}^{sg}_{c\_out}.$$

Spatial Attention Module



The input of *attributes*-guided branch:

$$\mathbf{F}^{ag}_{s\_inp} = [\mathbf{F}^{ag}_{c\_out}; \mathbf{A}] \in \mathbb{R}^{C' \times H \times W}$$

Attributes-guided spatial attention:

$$\mathbf{M}^{ag}_{s} = \sigma(f^{ag}([\mathrm{AvgPool}(\mathbf{F}^{ag}_{s\_inp}); \mathrm{MaxPool}(\mathbf{F}^{ag}_{s\_inp})])),$$

$$\mathbf{F}^{ag}_{s\_out} = \mathbf{M}^{ag}_{s} \otimes \mathbf{F}^{ag}_{c\_out},$$

The input of *self*-guided branch:

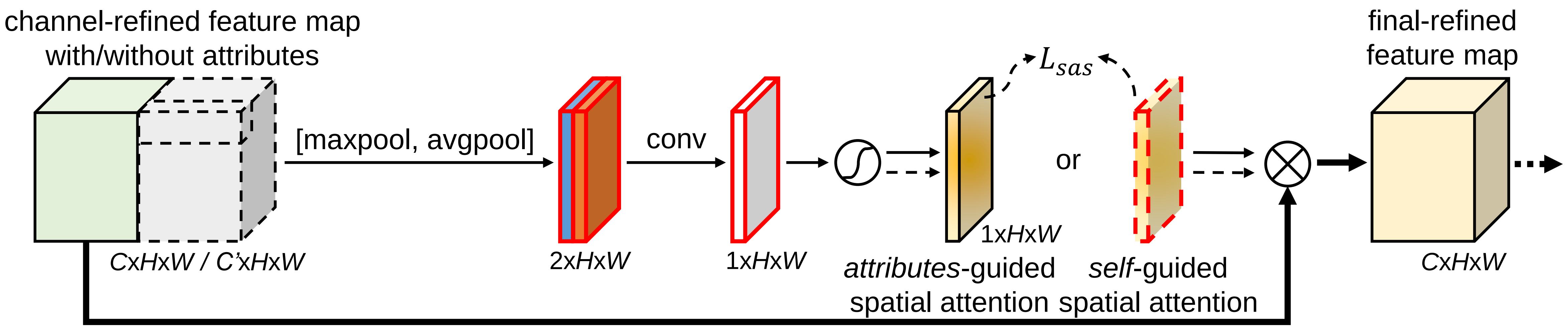
$$\mathbf{F}^{sg}_{s\_inp} = \mathbf{F}^{sg}_{c\_out} \in \mathbb{R}^{C \times H \times W}$$

Self-guided spatial attention:

$$\mathbf{M}^{sg}_{s} = \sigma(f^{sg}([\mathrm{AvgPool}(\mathbf{F}^{sg}_{s\_inp}); \mathrm{MaxPool}(\mathbf{F}^{sg}_{s\_inp})])),$$

$$\mathbf{F}^{sg}_{s\_out} = \mathbf{M}^{sg}_{s} \otimes \mathbf{F}^{sg}_{c\_out}.$$

Spatial Attention Module



The input of *attributes*-guided branch:

$$\mathbf{F}^{ag}_{s\_inp} = [\mathbf{F}^{ag}_{c\_out}; \mathbf{A}] \in \mathbb{R}^{C' \times H \times W}$$

Attributes-guided spatial attention:

$$\mathbf{M}^{ag}_{s} = \sigma(f^{ag}([\mathrm{AvgPool}(\mathbf{F}^{ag}_{s\_inp}); \mathrm{MaxPool}(\mathbf{F}^{ag}_{s\_inp})])),$$

$$\mathbf{F}^{ag}_{s\_out} = \mathbf{M}^{ag}_{s} \otimes \mathbf{F}^{ag}_{c\_out},$$

The input of *self*-guided branch:

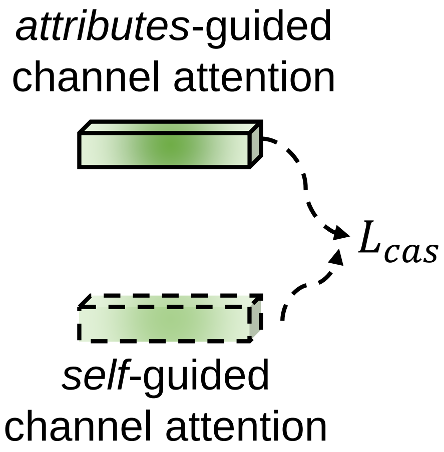
$$\mathbf{F}^{sg}_{s\_inp} = \mathbf{F}^{sg}_{c\_out} \in \mathbb{R}^{C \times H \times W}$$

Self-guided spatial attention:

$$\mathbf{M}^{sg}_{s} = \sigma(f^{sg}([\mathrm{AvgPool}(\mathbf{F}^{sg}_{s\_inp}); \mathrm{MaxPool}(\mathbf{F}^{sg}_{s\_inp})])),$$

$$\mathbf{F}^{sg}_{s\_out} = \mathbf{M}^{sg}_{s} \otimes \mathbf{F}^{sg}_{c\_out}.$$

Attention Alignment Mechanism



Channel attention alignment loss:

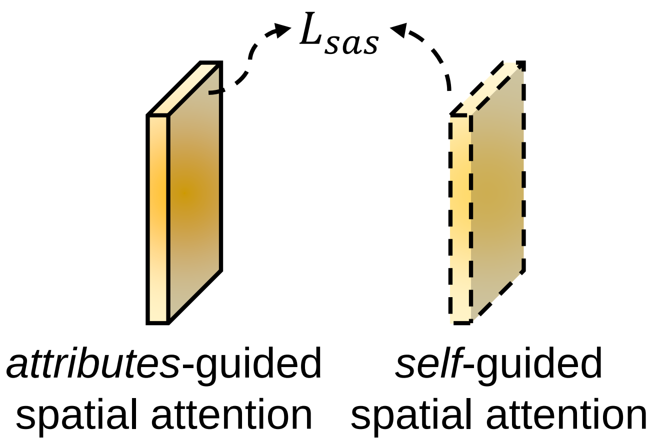
$$l_i^{cas} = \sum_j \log(1 + \exp(-\widetilde{\mathbf{M}}_c^{ag}(j) \otimes \widetilde{\mathbf{M}}_c^{sg}(j))),$$

$$L_{cas} = \sum_i^{N*K} l_i^{cas},$$

i : index of the support samples

$\widetilde{\mathbf{M}}$: normalized attention map

(j) : index of the element of the attention map



Spatial attention alignment loss:

$$l_i^{sas} = \sum_j \log(1 + \exp(-\widetilde{\mathbf{M}}_s^{ag}(j) \otimes \widetilde{\mathbf{M}}_s^{sg}(j))),$$

$$L_{sas} = \sum_i^{N*K} l_i^{sas}.$$

Overall loss function

Metric-based classification loss:

$$L_{mbc} = -\sum_{b=1}^{Q} \log p(y = y_n | v_b^q),$$

v_b^q : the feature embedding of the *b*-th query sample

$p(y = y_n | v_b^q)$: the probability of predicting the *b*-th query sample as the n-th class

The overall loss:

$$L = L_{mbc} + \alpha L_{cas} + \beta L_{sas}.$$

α, β : trade-off hyperparameters to balance the effects of different losses

Overall loss function

Metric-based classification loss:

$$L_{mbc} = -\sum_{b=1}^{Q} \log p(y = y_n | v_b^q),$$

v_b^q : the feature embedding of the *b*-th query sample

$p(y = y_n | v_b^q)$: the probability of predicting the *b*-th query sample as the n-th class

The overall loss:

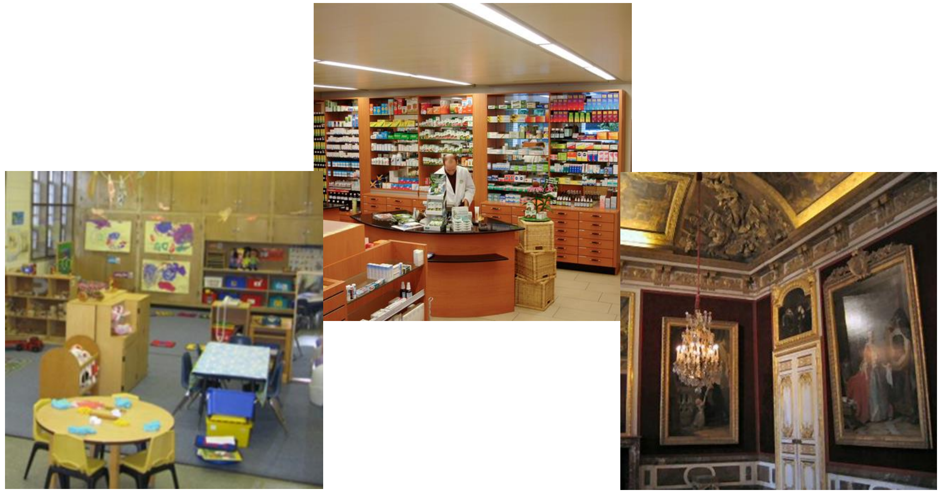
$$L = L_{mbc} + \alpha L_{cas} + \beta L_{sas}.$$

α, β : trade-off hyperparameters to balance the effects of different losses

Datasets

Caltech-UCSD-Birds 200-2011 (CUB)

- 11,788 images
- 200 categories: 100 / 50 / 50
- 312 category-level attributes



SUN Attribute Database (SUN)

- 14,340 images
- 717 categories: 580 / 65 / 72
- 102 image-level attributes

西湖大學 WESTLAKE UNIVERSITY

Main results

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

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

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

Main results

西湖大學 WESTLAKE UNIVERSITY

Method	Backbone	Test Accuracy 5-way 1-shot	5-way 5-shot
MatchingNet (Vinyals et al. 2016)	Conv-4	61.16 ± 0.89	72.86 ± 0.70
ProtoNet (Snell, Swersky, and Zemel 2017)	Conv-4	51.31 ± 0.91	70.77 ± 0.69
RelationNet (Sung et al. 2018)	Conv-4	62.45 ± 0.98	76.11 ± 0.69
MACO (Hilliard et al. 2018)	Conv-4	60.76	74.96
MAML (Finn, Abbeel, and Levine 2017)	Conv-4	55.92 ± 0.95	72.09 ± 0.76
Baseline (Chen et al. 2019a)	Conv-4	47.12 ± 0.74	64.16 ± 0.71
Baseline++ (Chen et al. 2019a)	Conv-4	60.53 ± 0.83	79.34 ± 0.61
Comp. (Tokmakov, Wang, and Hebert 2019) *	ResNet-10	53.6	74.6
AM3 (Xing et al. 2019) † *	Conv-4	73.78 ± 0.28	81.39 ± 0.26
AGAM (OURS) *	Conv-4	$\mathbf{75.87 \pm 0.29}$	$\mathbf{81.66 \pm 0.25}$
MatchingNet (Vinyals et al. 2016) †	ResNet-12	60.96 ± 0.35	77.31 ± 0.25
ProtoNet (Snell, Swersky, and Zemel 2017)	ResNet-12	68.8	76.4
RelationNet (Sung et al. 2018) †	ResNet-12	60.21 ± 0.35	80.18 ± 0.25
TADAM (Oreshkin, López, and Lacoste 2018)	ResNet-12	69.2	78.6
FEAT (Ye et al. 2020)	ResNet-12	68.87 ± 0.22	82.90 ± 0.15
MAML (Finn, Abbeel, and Levine 2017)	ResNet-18	69.96 ± 1.01	82.70 ± 0.65
Baseline (Chen et al. 2019a)	ResNet-18	65.51 ± 0.87	82.85 ± 0.55
Baseline++ (Chen et al. 2019a)	ResNet-18	67.02 ± 0.90	83.58 ± 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) $^{* \circ \bullet}$	DenseNet-121	76.1	82.9
Dual TriNet (Chen et al. 2019b) $^{* \circ}$	ResNet-18	69.61 ± 0.46	84.10 ± 0.35
AGAM (OURS) *	ResNet-12	$\mathbf{79.58 \pm 0.25}$	$\mathbf{87.17 \pm 0.23}$

Table 2: Average accuracy (%) comparison to state-of-the-arts with 95% confidence intervals on the CUB dataset. † denotes that it is our implementation. * denotes that it uses auxiliary attributes. $^{\circ}$ denotes that it uses auxiliary label embeddings. $^{\bullet}$ denotes that it uses auxiliary descriptions of the categories. Best results are displayed in **boldface**.

Ablation study

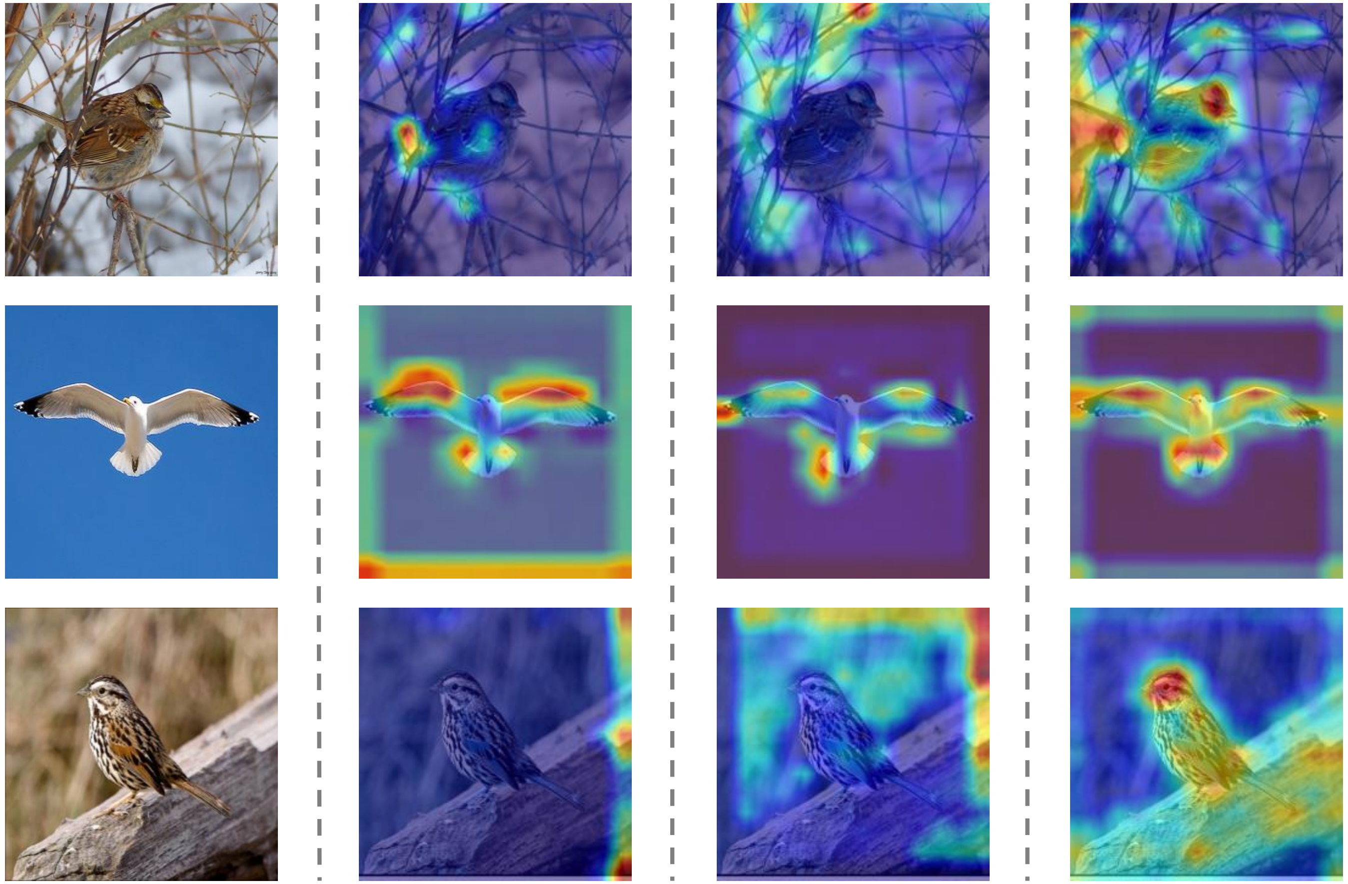
Loss Type	CUB	
	5-way 1-shot	5-way 5-shot
L1	66.95 ± 0.30	78.40 ± 0.25
MSE	69.83 ± 0.30	77.35 ± 0.22
smoothL1	72.42 ± 0.30	75.72 ± 0.31
soft margin	**75.87 ± 0.29**	**81.66 ± 0.25**
Loss Type	**SUN**	
	5-way 1-shot	5-way 5-shot
L1	60.56 ± 0.33	76.14 ± 0.26
MSE	59.54 ± 0.35	78.35 ± 0.26
smoothL1	62.07 ± 0.31	78.42 ± 0.23
soft margin	**65.15 ± 0.31**	**80.08 ± 0.21**

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	**75.87 ± 0.29**	**81.66 ± 0.25**
AGAM_SACA	74.22 ± 0.27	79.72 ± 0.26
w/o avgpool	66.27 ± 0.29	76.58 ± 0.25
w/o maxpool	67.60 ± 0.29	77.09 ± 0.22
w/o CA	54.91 ± 0.36	80.52 ± 0.24
w/o SA	69.66 ± 0.31	76.24 ± 0.27
w/o L_{cas}	74.88 ± 0.26	77.78 ± 0.26
w/o L_{sas}	74.29 ± 0.27	77.87 ± 0.23
w/o L_{cas}&L_{sas}	75.37 ± 0.31	78.92 ± 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**.

Attention visualization



Original images | ProtoNet | AGAM w/o attention alignment | Complete AGAM

Gradient-weighted class activation mapping (Grad-CAM) visualization of query samples.

Conclusion

- Using auxiliary semantic modalities in a proper manner contributes to few-shot recognition.

- We design similar feature selection processes for both support and query samples. When improving the discriminability with attributes-guided or self-guided channel and spatial attention, features extracted by both visual contents and corresponding attributes share the same space with pure-visual features.

Conclusion

- We propose an attention alignment mechanism between the attributes-guided and self-guided branches, so that the supervision signal from the attributes-guided branch promotes the self-guided branch to focus on more important features even without attributes.

- Extensive experiments show that our light-weight module can significantly improve metric-based approaches to achieve SOTA.

Thanks for watching

Siteng Huang

Min Zhang

Yachen Kang

Donglin Wang*

Project Page: https://kyonhuang.top/publication/attributes-guided-attention-module

Code: https://github.com/bighuang624/AGAM