All the attention you need: Global-local, spatial-channel attention for image retrieval

paper https://arxiv.org/abs/2107.08000

Chull Hwan Song, Odd Concepts

Hye Joo Han, Odd Concepts

Yannis Avrithis, Athena RC

Oddconcepts

ATHENA Research & Innovation Information Technologies

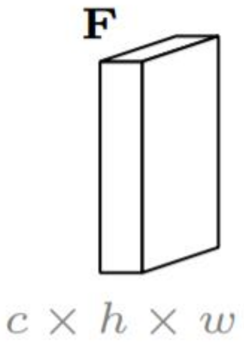
Introduction

- Goal: introduce a novel representation learning method for instance-level image retrieval
- Global-Local Attention Module (**GLAM**)
  - Attached at the end of backbone
  - All four forms of attention: either **local** or **global,** and either **spatial** or **channel**
- Contributions
  - First study employing all four forms of attention
  - Empirical evidence of the interaction of all forms of attention
  - State of the art on global descriptors (no re-ranking) for image retrieval

Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

Global-local attention module (GLAM)

local attention



local (1st order) attention:

- weigh channels and spatial locations **independently**

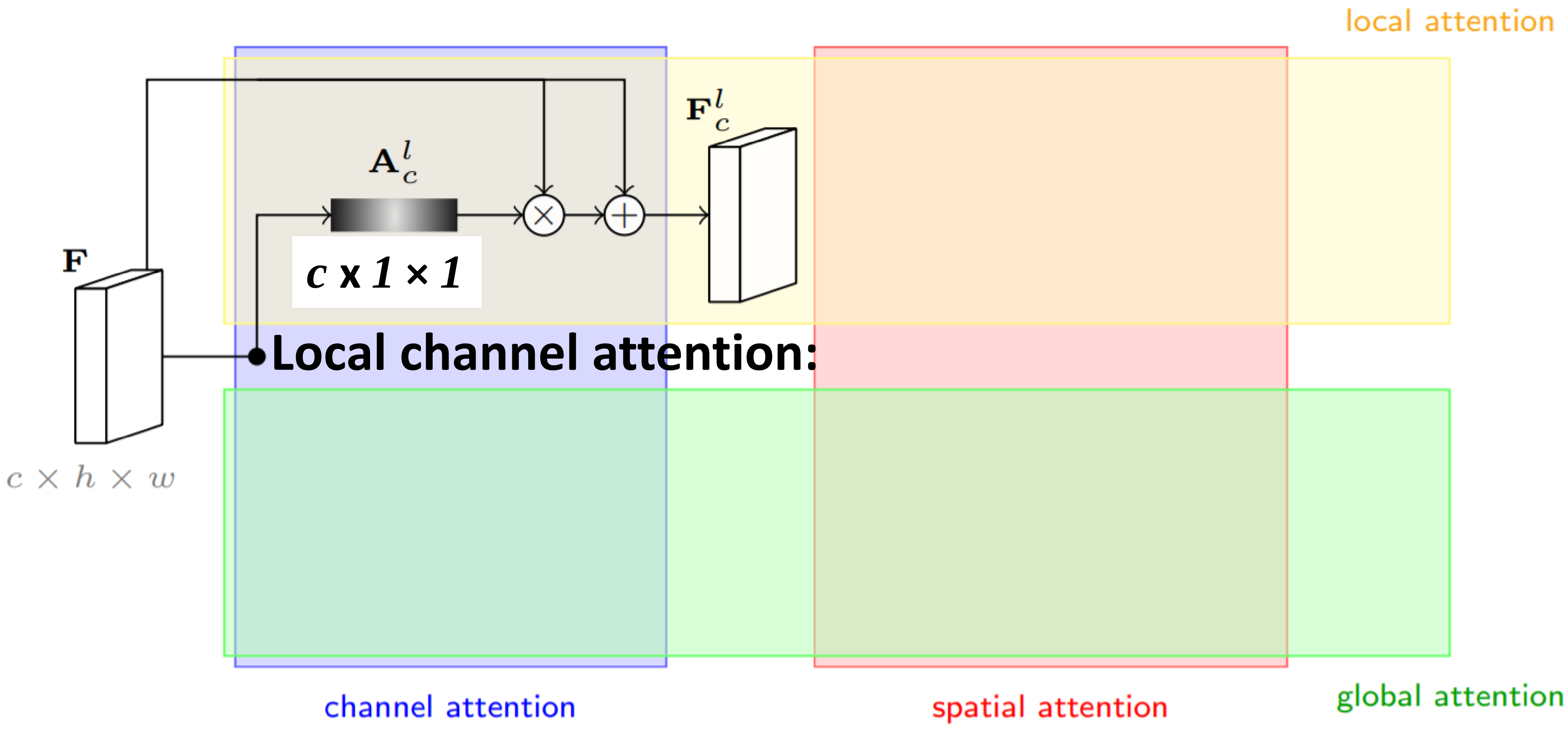
global (2nd order) attention:

- capture **pairwise interaction** within channels or within spatial locations

global attention

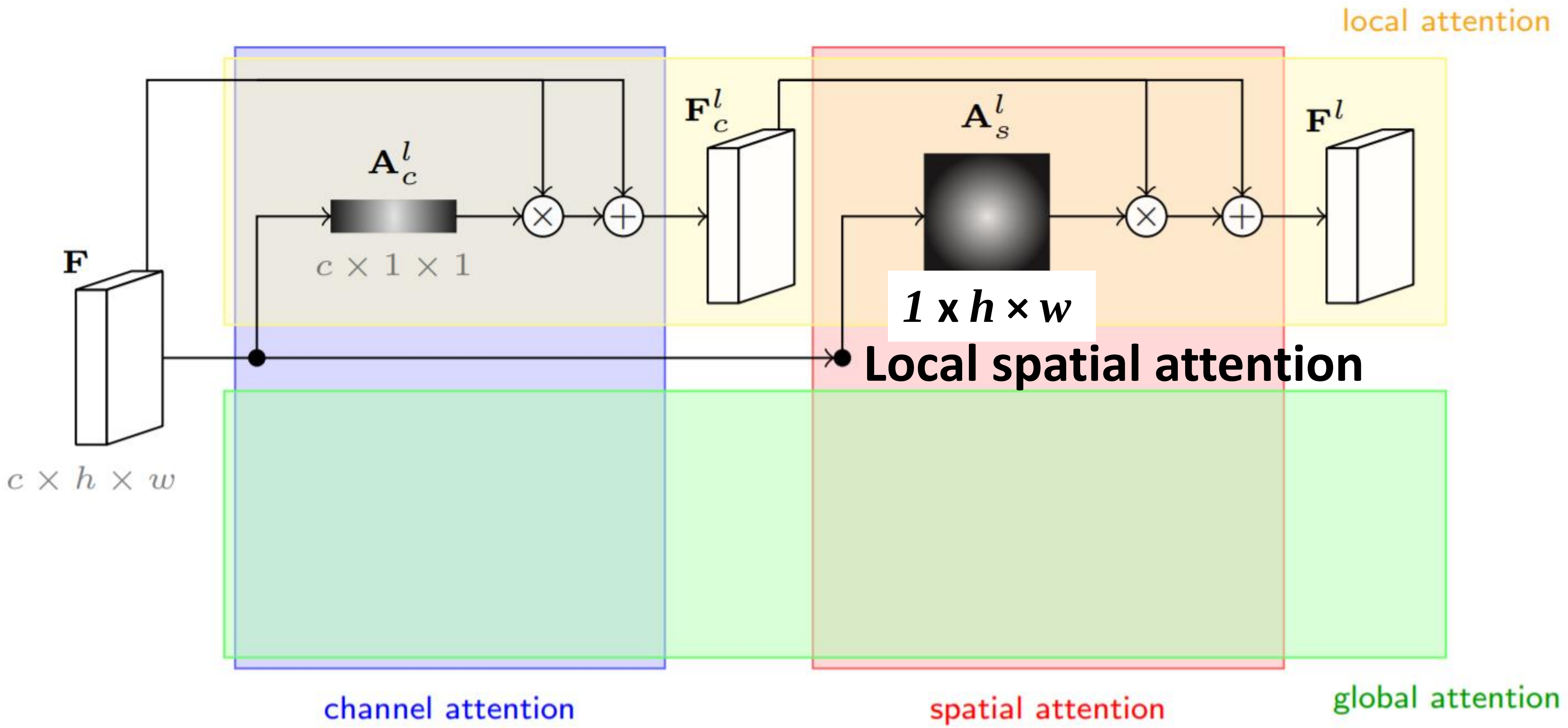
- Collect contextual information from the feature tensor F through **two parallel** network streams, **local** and **global** attention

Global-local attention module (GLAM)



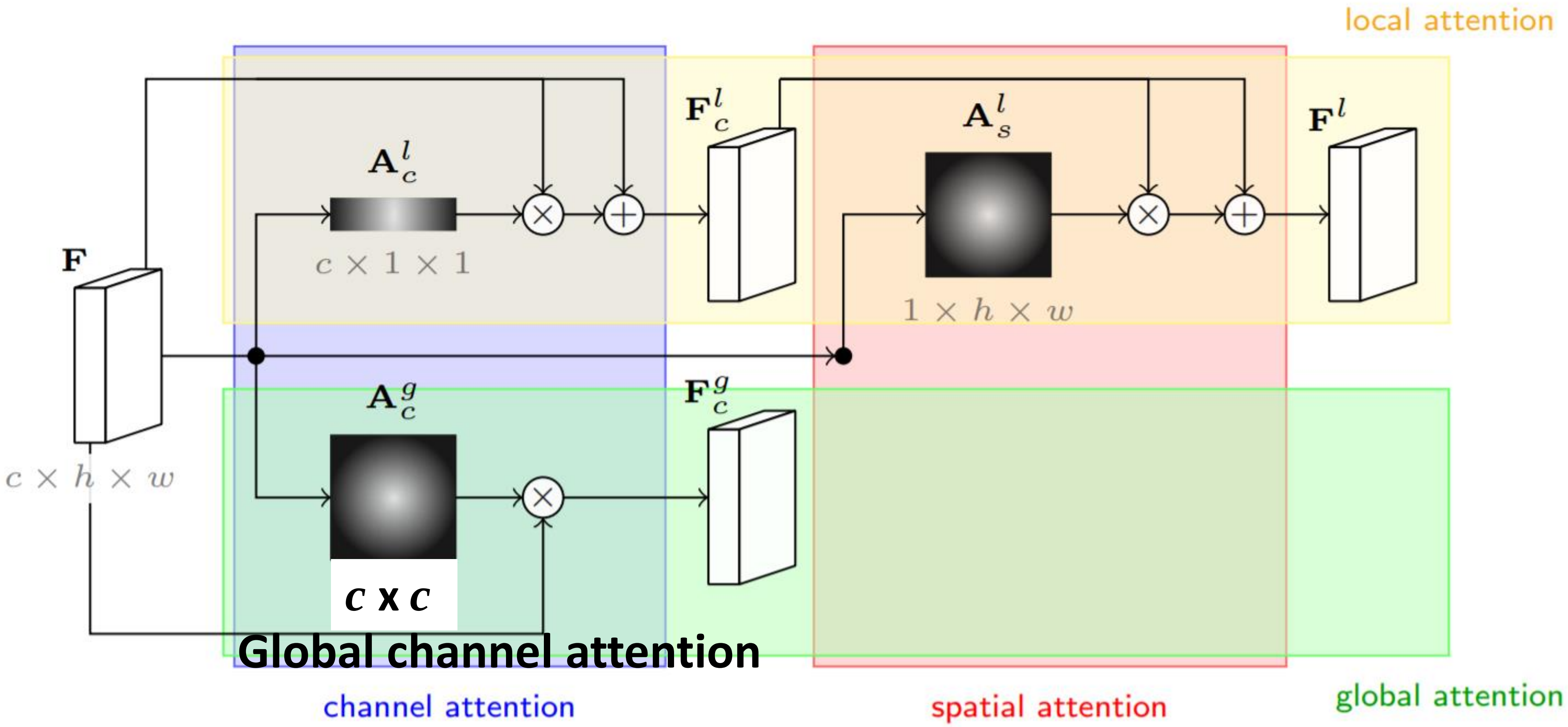
- GAP + conv1d yields ***c* x *1* x *1*** local channel attention maps

Global-local attention module (GLAM)



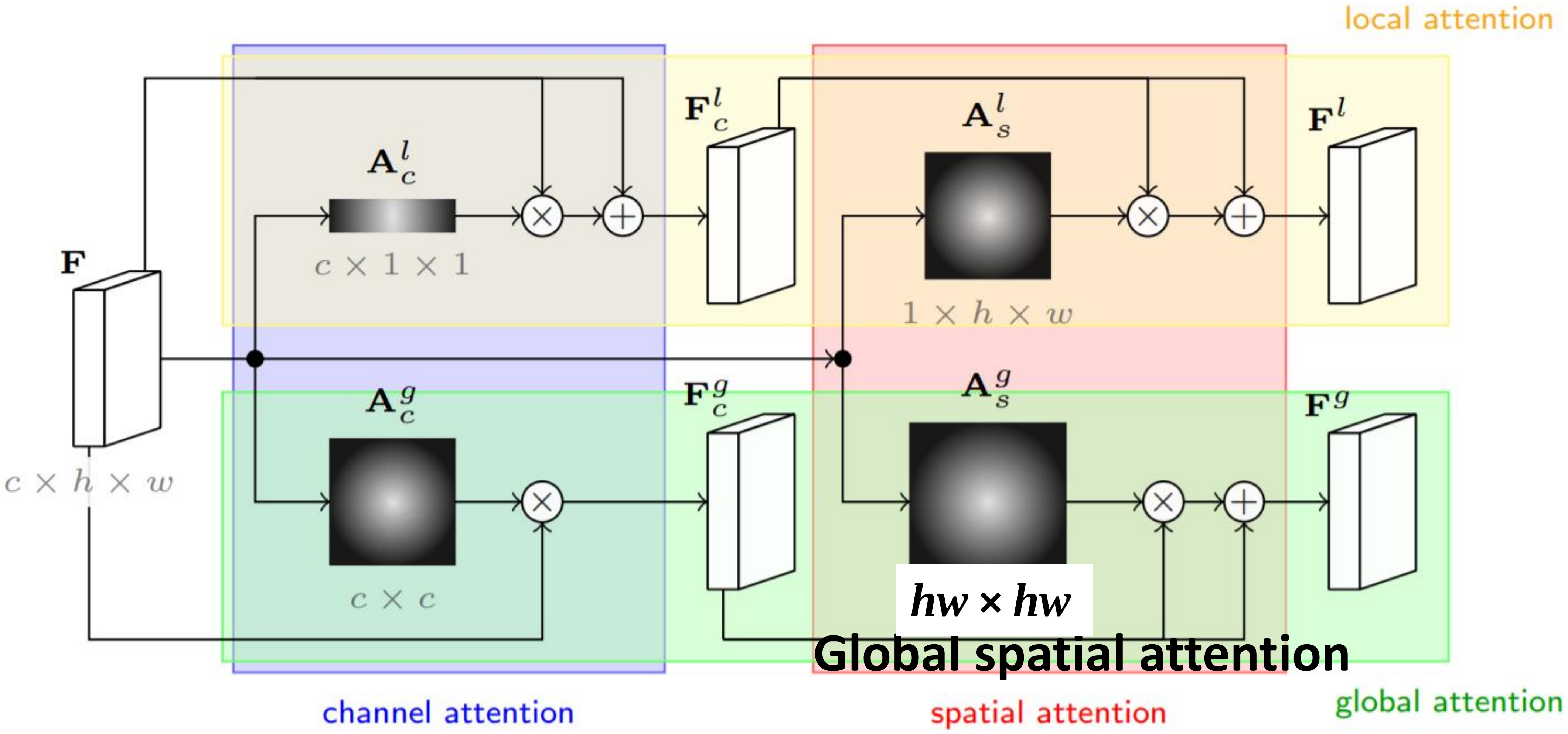
- Conv layers yield $1 \times h \times w$ local spatial attention maps

Global-local attention module (GLAM)



- GAP + conv1d yields c x c global channel attention maps

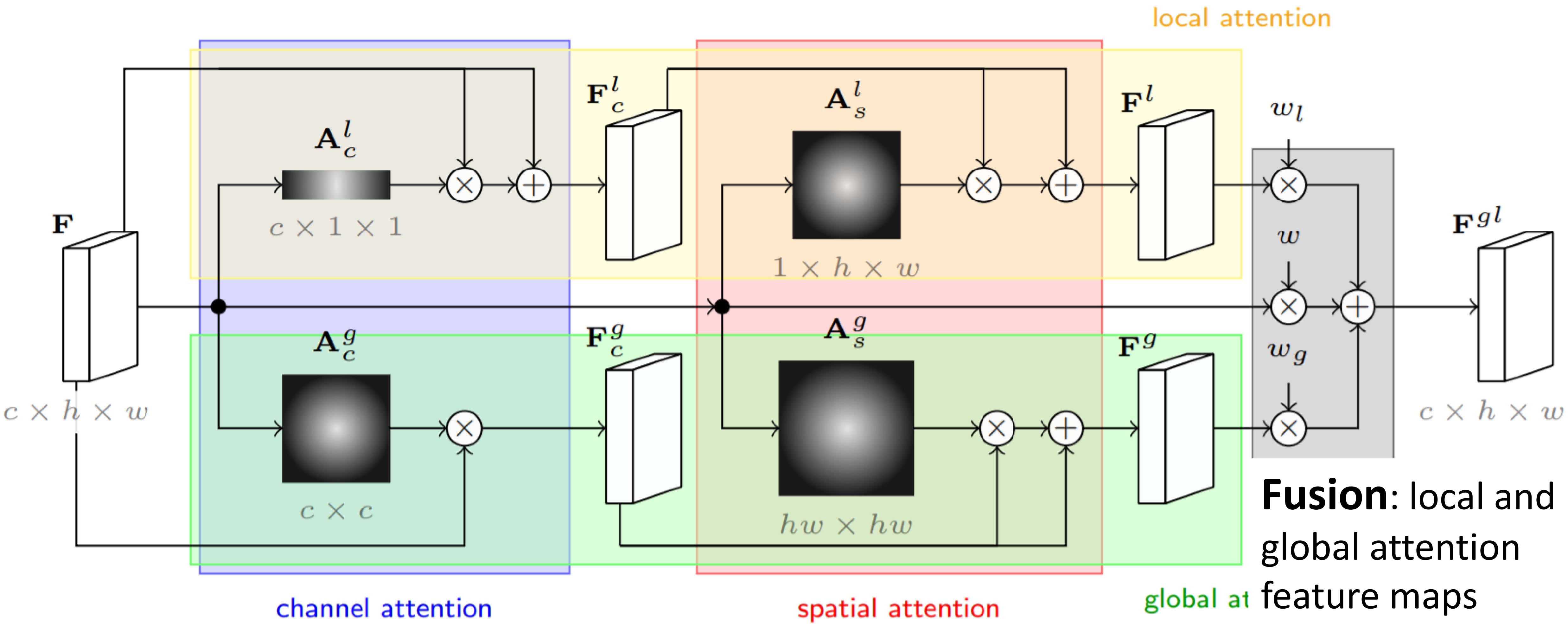
Global-local attention module (GLAM)



- Conv layers yield ***hw* x *hw*** global spatial attention maps

Song, Han and Avrithis. WACV 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval.

Global-local attention module (GLAM)



Fusion: local and global attention feature maps fused with the original feature

Datasets and implementation details

- ResNet101-GeM pooling
- Final embedding: 512 dimension
- Global descriptor only, without re-ranking
- Test set: Oxford5k, Paris6k, Revisited Oxford ($\mathcal{R}$Oxf)/Paris ($\mathcal{R}$Par)
- Metrics: mean average precision (mAP)

State of the art comparisons

METHOD	TRAIN SET	DIM	OXF5K	PAR6K	$\mathcal{R}$MEDIUM		$\mathcal{R}$HARD	
					$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GeM-Siamese [37, 35]	SfM-120k	2048	87.8	92.7	64.7	77.2	38.5	56.3
SOLAR [28]	GLDv1-noisy	2048	–	–	69.9	81.6	47.9	64.5
DELG [5]	GLDv1-noisy	2048	–	–	73.2	82.4	51.2	64.7
GLDv2 [53] (Weyand)	GLDv2-clean	2048	–	–	74.2	84.9	51.6	70.3
GLAM (Ours)	NC-clean	512	77.8	85.8	51.6	68.1	20.9	44.7
	GLDv1-noisy	512	92.8	95.0	**73.7**	**83.5**	49.8	**69.4**
	GLDv2-noisy	512	93.3	95.3	75.7	86.0	53.1	73.8
	GLDv2-clean	512	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

All use ResNet101-GeM. Red: best results. Blue: GLAM higher than DELG on GLDv1-noisy

State of the art comparisons

METHOD	TRAIN SET	DIM	OXF5K	PAR6K	$\mathcal{R}$MEDIUM		$\mathcal{R}$HARD	
					$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GeM-Siamese [37, 35]	SfM-120k	2048	87.8	92.7	64.7	77.2	38.5	56.3
SOLAR [28]	GLDv1-noisy	2048	–	–	69.9	81.6	47.9	64.5
DELG [5]	GLDv1-noisy	2048	–	–	73.2	82.4	51.2	64.7
GLDv2 [53] (Weyand)	GLDv2-clean	2048	–	–	74.2	84.9	51.6	70.3
GLAM (Ours)	NC-clean	512	77.8	85.8	51.6	68.1	20.9	44.7
	GLDv1-noisy	512	92.8	95.0	**73.7**	**83.5**	49.8	**69.4**
	GLDv2-noisy	512	93.3	95.3	75.7	86.0	53.1	73.8
	GLDv2-clean	512	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

All use ResNet101-GeM. Red: best results. Blue: GLAM higher than DELG on GLDv1-noisy

Effect of attention modules

Method	Oxf5k	Par6k	$\mathcal{R}$Medium		$\mathcal{R}$Hard	
			$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GLAM baseline	91.9	94.5	72.8	84.2	49.9	69.7
+local-channel	91.3	95.3	72.2	85.8	48.3	73.1
+local-spatial	91.0	95.1	72.1	85.3	48.3	71.9
+local	91.2	95.4	73.7	86.5	52.6	75.0
+global-channel	92.5	94.4	73.3	84.4	49.8	70.1
+global-spatial	92.4	95.1	73.2	86.3	50.0	72.7
+global	92.3	95.3	77.2	86.7	57.4	75.0
+global+local	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

Local channel/spatial attention:
- Sometimes harmful when used alone
- But beneficial when used together (+local)

Effect of attention modules

METHOD	OXF5K	PAR6K	$\mathcal{R}$MEDIUM		$\mathcal{R}$HARD	
			$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GLAM baseline	91.9	94.5	72.8	84.2	49.9	69.7
+local-channel	91.3	95.3	72.2	85.8	48.3	73.1
+local-spatial	91.0	95.1	72.1	85.3	48.3	71.9
+local	91.2	95.4	73.7	86.5	52.6	75.0
+global-channel	92.5	94.4	73.3	84.4	49.8	70.1
+global-spatial	92.4	95.1	73.2	86.3	50.0	72.7
+global	92.3	95.3	77.2	86.7	57.4	75.0
+global+local	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

Global channel/spatial attention:
- mostly beneficial even when used alone
- Impressive gain when used together (+global)

Effect of attention modules

Method	Oxf5k	Par6k	$\mathcal{R}$Medium		$\mathcal{R}$Hard	
			$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GLAM baseline	91.9	94.5	72.8	84.2	49.9	69.7
+local-channel	91.3	95.3	72.2	85.8	48.3	73.1
+local-spatial	91.0	95.1	72.1	85.3	48.3	71.9
+local	91.2	95.4	73.7	86.5	52.6	75.0
+global-channel	92.5	94.4	73.3	84.4	49.8	70.1
+global-spatial	92.4	95.1	73.2	86.3	50.0	72.7
+global	92.3	95.3	77.2	86.7	57.4	75.0
+global+local	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

global+local attention:
- Further improvement
- Shows necessity of both attention

Conclusions

- Novel approach for extracting global and local contextual information using attention mechanisms operating on both spatial and channel dimensions
- Comprehensive study and empirical evaluation of all four forms of attention for instance-level image retrieval
- Maximum gain when all forms are present

Thank you!