Supplimentary: GAF-Net: Improving the Performance of Remote Sensing Image Fusion using Novel Global Self and Cross Attention Learning

Ankit Jha[1*] Shirsha Bose[2*] Biplab Banerjee[1]
[1]Indian Institute of Technology Bombay, India [2]Technical University of Munich, Germany
{ankitjha16, shirshabosecs, getbiplab}@gmail.com

1. Introduction

In the supplementary paper, we provide the following analysis:

a. Architecture ablation: We present the training curves of GAF-Net on the Houston HSI-LiDAR dataset in Figure 1 and ADVANCE dataset in Figure 2 and justify the convergence of FAG-Net over the epochs.

b. The analysis for the baseline architectures and self-attention block (SAB) on the ADVANCE dataset, as shown in Table 1. We incrementally consider all the model components (SAB, $\mathcal{L}_{NRR}$ and CAB). It can be clearly observed that the application of these components progressively improves the performance.

c. **Analysis of $Fe_1^{\mathcal{S}}$ and $Fe_2^{\mathcal{S}}$:** Tables 2 and 3 show the analysis on the depth of Fe_1^{S} and Fe_2^{S} for fixed $Fe^{\mathcal{T}_1}$ and $Fe^{\mathcal{T}_2}$ (ResNet-50 [3]) in GAF-Net (residual blocks with SAB) on both HSIs and audio-visual datatets. By considering all the multimodal dataset used in our experiments, we found that typically, $\mathcal{S}$ with four conv. blocks for each of the streams provide consistently superior performance.

d. **Analysis of CAB:** In Table 4, we showcase the importance of cross-attention generation from the deepest layers of $Fe^{\mathcal{T}_1}$ and $Fe^{\mathcal{T}_2}$, we perform the following experiments, i) GAF-Net without CAB, ii) cross-attention generated from the intermediate layers of $Fe^{\mathcal{T}_1}$ and $Fe^{\mathcal{T}_2}$, and iii) cross-attention generated from the intermediate self-attended outputs $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{X}}_2$, respectively. We observe from Table 4 that the proposed CAB outperforms the remaining baselines significantly, at least by 1.8% in precision, recall and F1 values.

e. **Sensitivity to the amount of training samples and effect of our proposed attention modules:** In order to analyse the agnostic property of our proposed modules, we have ablate the model without and with attention on two different backbones (DenseNet-121 and ResNet-18). We observe that the proposed attention modules help in better classification performance on both HSIs (in Figure 3) and ADVANCE datasets (in Figure 4).

f. **Sensitivity to the amount of training samples and effect of different self-attention modules for ADVANCE dataset:** We present the bar graphs for different ablations, i.e., Figure 5 (a) and (b) represent the sensitivity to the amount of training samples and comparison on the attention modules, respectively.

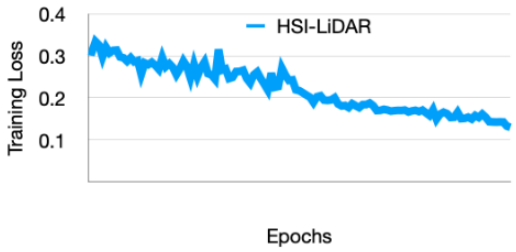
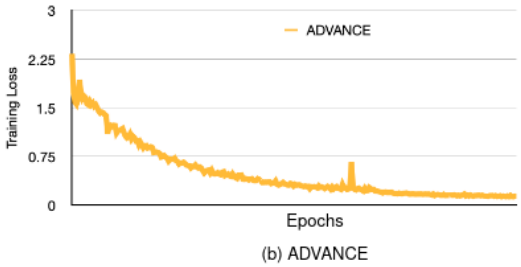


Figure 1. The training curve for GAF-Net on the Houston HSI-LiDAR datasets.



(b) ADVANCE

Figure 2. The training curve for GAF-Net on the ADVANCE datasets.

Ablation on pairs of bi-modality in the Augsburg dataset: In Table 5, we present the ablations on modality

*equal contribution

Table 1. Ablation analysis of our proposed GAF-Net on the ADVANCE dataset. **A** and **B** are the defined baselines and analysis on SAB, respectively. † represents only the sub-network $\mathcal{S}$ and without CAB. SA, GCA, and LCA represent spatial attention 1, global channel attention, and local channel attention, respectively. MSFRB is multi-scale feature refinement block. We highlighted the best results in **bold**.

Methods					ADVANCE		
A: Baselines					Precision	Recall	F1
	Layer-wise SAB	CAB	$\mathcal{L}_{NRR}$	MSRFB			
A1:	✗	✗	✗	✗	87.63	88.10	87.86
A2:	✗	✗	✓	✗	88.93	89.24	89.08
A3:	✓	✗	✗	✗	88.76	89.12	88.93
A4:	✓	✗	✓	✗	89.18	89.21	89.19
A5:	✗	✓	✗	✗	89.04	88.93	88.98
A6:	✗	✗	✗	✓	89.17	88.99	89.07
A7:	✗	✓	✓	✗	91.11	90.97	91.04
A8:	✗	✓	✓	✓	91.95	92.11	92.03
A9:	✓	✓	✗	✓	92.43	92.24	92.33
A10:	✓	✓	✓	✗	92.82	92.73	92.77
B: Ablation on SAB							
B1:	SA				91.69	91.56	91.62
B2:	GCA				91.54	91.58	91.56
B3:	LCA				90.02	89.80	89.91
B4:	SA + GCA				92.32	92.45	92.38
B5:	SA + LCA				92.09	92.14	92.15
B6:	SA + GCA + LCA †				92.46	92.78	92.61
GAF-Net					**93.37**	**93.23**	**93.31**

Table 2. The depth analysis for the streams of $Fe_1^{\mathcal{S}}$ and $Fe_2^{\mathcal{S}}$ in GAF-Net (residual blocks with SAB) on Houston 2013 HSI-LiDAR and HSI-MSI, Berlin HSI-SAR, and Augsburg HSI-SAR hyperspectral datasets. N denotes number of blocks in $\mathcal{S}$. We highlighted the best results in **bold**.

N-blocks in S	Houston2013 HSI-LiDAR			Houston2013 HSI-MSI			Berlin HSI-SAR			Augsburg HSI-SAR		
	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ
2-blocks	90.71	93.64	0.8950	90.02	92.78	0.8893	77.95	69.93	0.6701	89.22	68.87	0.8432
4-blocks	91.39	**94.92**	0.9018	**90.64**	93.30	**0.8938**	78.57	**70.92**	**0.6761**	**90.80**	70.10	**0.8683**
6-blocks	**91.45**	94.38	**0.9039**	90.33	92.97	0.8911	**78.70**	70.33	0.6750	89.76	69.10	0.8544
8-blocks	91.23	94.34	0.8996	90.59	**93.50**	0.8925	78.20	70.52	0.6722	90.53	**70.12**	0.8631

Table 3. The depth analysis for the streams of $Fe_1^{\mathcal{S}}$ and $Fe_2^{\mathcal{S}}$ in GAF-Net (residual blocks with SAB) on the ADVANCE dataset. N denotes number of blocks in $\mathcal{S}$. We highlighted the best results in **bold**.

N-blocks in S	ADVANCE		
	Precision	Recall	F1
2-blocks	92.89	92.97	92.93
4-blocks	93.37	93.23	**93.31**
6-blocks	**93.46**	93.04	93.25
8-blocks	93.16	**93.41**	93.28

Table 4. Ablation analysis on CAB for the ADVANCE dataset. IL in * and FL in ** denote Intermediate Layer and Final Layer, respectively. We highlighted the best results in **bold**.

Methods	ADVANCE		
CAB Ablation	Precision	Recall	F1
No CAB	89.15	89.38	89.26
CAB from IL of $\mathcal{T}_1$ and $\mathcal{T}_2$ *	91.65	91.59	91.62
CAB from FL of $\mathcal{S}$ **	90.89	90.91	90.90
GAF-Net	**93.37**	**93.23**	**93.31**

combination for the hyperspectral (HSI) Augsburg dataset and choose to work with HSI and synthetic aperture radar (SAR) as the multimodal data. Our GAF-Net outperforms

Table 5. The ablation results on modality combination using GAF-Net for Augsburg dataset. $^{+}$ setting used in comparison with the other SOTA methods. We highlighted the best results in **bold**.

Modality Combination For Augsburg Dataset	OA
Combination of HSI + SAR^{+}	**90.80**
Combination of HSI + DSM	86.62
Combination of DSM + SAR	84.18

the other combinations, i.e., HSI-DSM and DSM-SAR at least by 3 %.

Generated classification maps for Houston2013 HSI-MSI dataset: Finally, in Figure 6, we show the classification maps between various SOTA methods and our proposed GAF-Net for Houston2013 HSI-MSI dataset, observe highly mapping of the predicted classes from GAF-Net with the ground truth.

Series and parallel combination of SA, GCA and LCA: Figure 7 justifies the use of SA, GCA, and LCA in SAB in our proposed architecture, i.e., we use the parallel combination of SA, GCA, and LCA in the SAB (as mentioned in

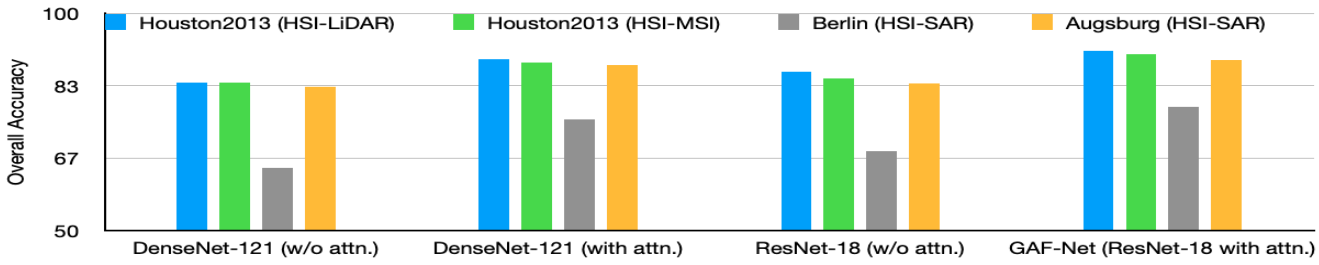


Figure 3. The performance comparison for our proposed attention modules on different backbones i.e. DenseNet-121 and ResNet-18 on Houston (HSI-LiDAR), Berlin (HSI-SAR), and Augsburg (HSI-SAR) datasets. Here, w/o attn. and with attn. denote the architecture without and with attention modules.

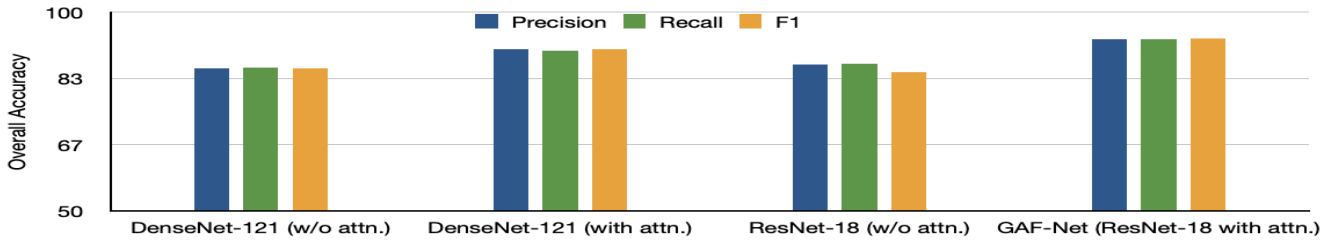


Figure 4. The performance comparison for our proposed attention modules on different backbones i.e. DenseNet-121 and ResNet-18 on the ADVANCE dataset. Here, w/o attn. and with attn. denote the architecture without and with attention modules.

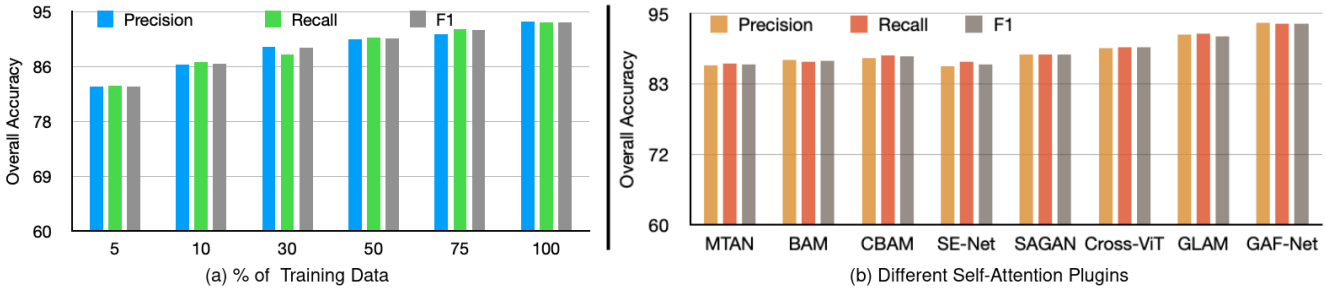


Figure 5. Analysis on (a) % of training data used in training, (b) feature extractors for $\mathcal{S}$, and (c) different self-attention plugins for our proposed GAF-Net architecture on ADVANCE dataset.

Figure 3(d) of main paper. We kept the position of CAB similar to which we have mentioned in Figure 2 (main paper). We present the bar graphs for all the the multimodal datasets (HSI and ADVANCE).

Class activation map (CAM) visualization and effect of learning the auxiliary network $\mathcal{T}_1$ and $\mathcal{T}_2$ from pre-training and scratch: Figure 8 shows the class activation maps (CAM) in highlighting the importance of our proposed GAF-Net. We also ablate between freeze (pre-trained) networks and training GAF-Net from scratch, shown in Figure 9.

Computation complexity and number of trainable attention parameters: We perform our experiments using Pytorch with 16 GB NVIDIA 3080 Ti GPU, 64 GB RAM and Intel Xeon processor. We analyse on the number of trainable parameters used by different SOTA attention modules and our GAF-Net, shown in Table 6. Here, the number of attention parameters in GAF-Net are lesser than CBAM [7], BAM [6], SE-Net [4], ViT [2] and cross-ViT [1], but comparable to MTAN [5].

Table 6. The depth analysis for the streams of $Fe_1^{\mathcal{S}}$ and $Fe_2^{\mathcal{S}}$ in GAF-Net (residual blocks with SAB) on the ADVANCE dataset. N denotes number of blocks in $\mathcal{S}$. We highlighted the best results in **bold**.

Attention Module	CBAM	BAM	SE-Net	MTAN	ViT	Cross-ViT	GAF-Net
Number of in Millions (M))	1.18	2.20	4.70	0.59	7.08	3.74	0.61

References

[1] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *CoRR*, abs/2103.14899, 2021.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

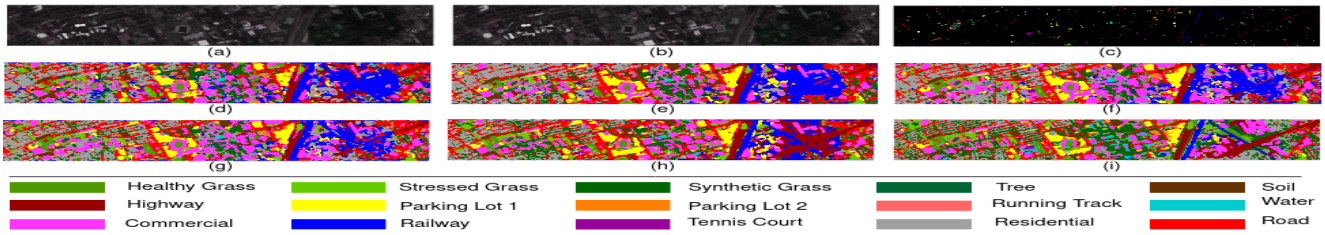


Figure 6. The classification maps for Houston2013 HSI-MSI-dataset. (a) HSI's true colour composite (RGB Bands-43, 22, 36), (b) MSI's true color composite (RGB Bands-2,1,4), and (c) Ground-truth. From (c) to (i) represent classification maps for different methods/algorithms. Specifically, (d)End-Net, (e)Co-CNN, (f) CCR-Net, (g) FusAtNet, (h) $\mathcal{T}_1 + \mathcal{T}_2$, (i) GAF-Net

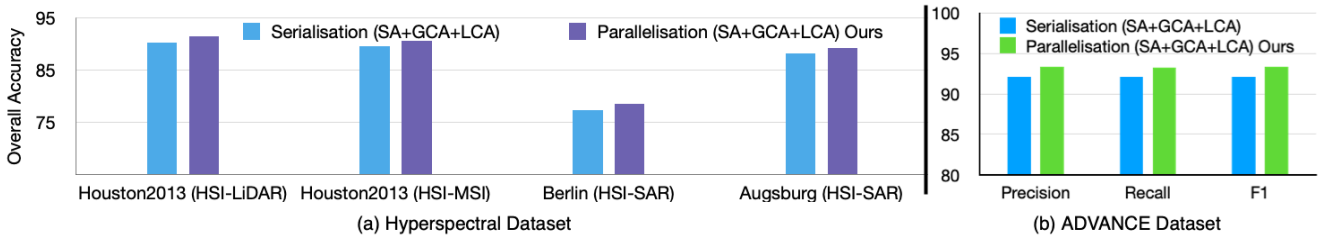


Figure 7. Analysis on series and parallel combinations of SA, GCA and LCA in SAB for (a) Hyperspectral Datasets and (b) ADVANCE dataset. We kept CAB to be at the similar position as mentioned for GAF-Net (main paper).

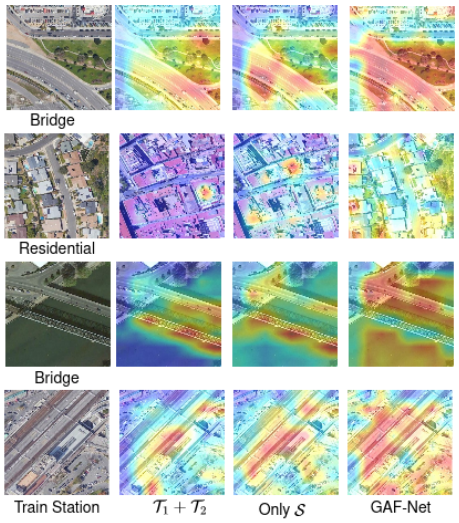


Figure 8. The generated class activation map from the fusion modality-specific network $\mathcal{T}_1$ and $\mathcal{T}_2$, i.e., $\mathcal{T}_1 + \mathcal{T}_2$, sub-network $\mathcal{S}$ (only $\mathcal{S}$ represents S with SAB no CAB) and GAF-Net on ADVANCE dataset.

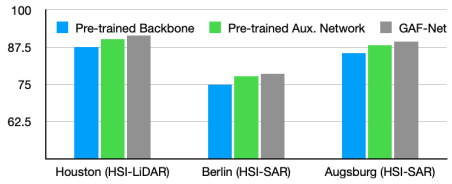


Figure 9. The performance comparison of GAF-Net with pre-trained backbone, pre-trained modality on Houston (HSI-LiDAR), Berlin (HSI-SAR), and Augsburg (HSI-SAR) datasets. Here, the pre-trained backbone and modality represent initializing the network with pre-trained weights on ImageNet and the datasets used in our experiments.

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[5] Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. *CoRR*, abs/1803.10704, 2018.

[6] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. BAM: bottleneck attention module. *CoRR*, abs/1807.06514, 2018.

[7] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.