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

Abstract

*The notion of self and cross-attention learning has been found to substantially boost the performance of remote sensing (RS) image fusion. However, while the self-attention models fail to incorporate the global context due to the limited size of the receptive fields, cross-attention learning may generate ambiguous features as the feature extractors for all the modalities are jointly trained. This results in the generation of redundant multi-modal features, thus limiting the fusion performance. To address these issues, we propose a novel fusion architecture called Global Attention based Fusion Network (GAF-Net), equipped with novel self and cross-attention learning techniques. We introduce the within-modality feature refinement module through **global** spectral-spatial attention learning using the query-key-value processing where both the global spatial and channel contexts are used to generate two channel attention masks. Since it is non-trivial to generate the cross-attention from within the fusion network, we propose to leverage two **auxiliary** tasks of modality-specific classification to produce highly discriminative cross-attention masks. Finally, to ensure non-redundancy, we propose to penalize the high correlation between attended modality-specific features. Our extensive experiments on five benchmark datasets, including optical, multispectral (MS), hyperspectral (HSI), light detection and ranging (LiDAR), synthetic aperture radar (SAR), and audio modalities establish the superiority of GAF-Net concerning the literature.*

1. Introduction

Recent times have witnessed the rapid development of remote sensing (RS) imaging techniques for precisely monitoring the Earth's surface. These images have direct applications in urban planning, environmental monitoring, geology, etc. [42, 1]. Amongst different RS data modalities, hyperspectral images (HSI) is characterized by practically continuous spectral properties, while the multispectral images (MSI) can provide finer spatial information. On the other hand, the SAR data or the elevation data generated from LiDAR are agnostic to atmospheric perturbations. In parallel to these visual modalities, audio is regarded as an important source of information for recognizing certain phenomenon, particularly in defense applications like military speech intelligence detection, military target detection, and disaster management, where it may be difficult to recognize some phenomenon from low-quality image feeds but can be identified using respective sound primitives. A few endeavors have explored the possibilities of synergistically fusing RS visual data with audio data [20, 16]. Multiple data sources, if combined intelligently, are able to produce discriminative and semantically rich features, something the individual modalities may not be capable of achieving.

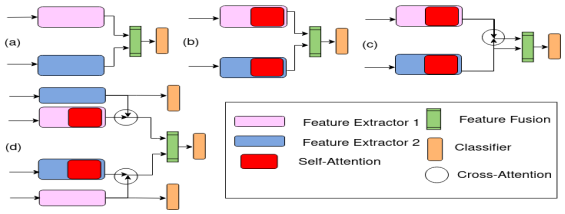

Figure 1. The evolution of fusion networks in RS. (a) standard feature extraction cum pooling based network, (b) each feature extractor has self-attention blocks before pooling is performed, (c), network with self and cross attention, (d) ours GAF-Net.

The multi-stream CNN-based deep learning models are predominantly utilized for unifying the feature information from multiple modalities into a combined representation space [51, 12, 10, 19, 49, 45]. In order to highlight important modality-specific features while suppressing any irrelevant information, the notion of *self-attention* learning while *disentangling the spatial and spectral components* has subsequently been introduced within the CNN framework [31, 48, 36, 22]. However, the CNN-based fusion networks coupled with self-attention do not interact with the cross-modal feature extractors. This causes some important shared high-level features from all the modalities to be overlooked. In addition, such a paradigm may make different features significantly unbalanced where each of the

*equal contribution

modality-specific feature may not be equally discriminative [33, 47]. The idea of *cross-attention* is envisioned as a remedy where the feature extractors of different modalities can influence each other (Figure 1).

One major problem of the existing self-attention learning techniques is that they are based on localized convolution operations. It suggests that the effects of only the neighboring pixels are counted while learning the attention mask for a given pixel. This is valid for the existing spatial and spectral attention learning modules. Notwithstanding the above, the global contextual information is found to boost the performance of the dense prediction tasks like land-cover classification (for example, the building the road pixels should be cooccuring within a context), as suggested in the literature [34]. Global feature learning can deal with the fragmentation problem of local models [34]. In this regard, the multi-head attention of transformers [43, 6, 29] implements global spatial attention by assessing the pairwise similarities among the image patches, though they are not designed to take care of the channel attention explicitly. This leads to the first research question we ask: *how to learn disentangled global spectral-spatial self-attention masks?*

Similarly, we argue that many cross-attention learning techniques [33] are not well calibrated. This is because these masks are learned from the individual modality-specific streams in parallel without being concerned about their discriminative nature. As a result, such cross-attention may limit the generalization ability of the fused features by injecting redundancy or highlighting ambiguous cross-modal features. This opens up the avenue for the important research direction on *how to learn discriminative and high-level cross-attention masks without affecting the feature learning of the fusion network?*

Finally, we must enforce the non-redundancy of the features before fusing them to avoid overfitting. The application of the attention modules ensures good modality-specific feature learning but does not explicitly ensures non-redundancy between them. This leads to our final research agenda of *how to penalize high correlation between the modality-specific features for fusion?*

Contributions: To solve the abovementioned issues, we propose a generic feature fusion network called GAF-Net for land-cover classification from bi-modal RS images. GAF-Net considers a global attention learning strategy for feature refinement by removing redundant and irrelevant information. We propose novel disentangled spectral-spatial self-attention and cross-modal attention learning to aid in better modality-specific feature learning. While we re-engineer the spatial attention module of transformers by using the residual connection for better multi-scale information propagation, we propose two novel channel attention modules for capturing the local and global variations of the spectral signatures for the classes.

For generating the cross-attention masks, we hypothesize that the discriminative and high-level feature embeddings of a given modality should be considered to generate the cross-attention masks. This is non-trivial to obtain from the fusion network as all the modality-specific streams are jointly trained. We propose supplementing the fusion network with two auxiliary modality-specific classification networks as a remedy. The cross-attention masks are generated from the deepest feature layers of these networks, which are highly discriminative and semantically superior. Such high-level cross-modal information helps express important hidden patterns from the modality features. Finally, we introduce a novel non-redundancy regularizer on the attended (application of the self followed by cross attention masks) feature representations per modality which seeks to decorrelate them. We highlight the significant contributions as follows,

- We design a simple and generic bi-modal fusion network for RS data called GAF-Net to learn discriminative and compact features through novel attention learning-based feature refinement in a principled manner.
- To our knowledge, we propose the first non-local spectral-spatial self-attention learning module using key-value processing. Besides, we introduce the novel paradigm of cross-attention learning from auxiliary tasks. Finally, we propose explicitly reducing redundancy between the modality features through a novel regularizer.
- We compare our attention modules with existing counterparts on a variety of datasets (visual, audio, and depth modalities), showing that the proposed global self-attention convincingly beats the models based on local operations (see Figure 5 (c)). Similarly, we highlight the superiority of proposed cross attention through extensive ablations. We strongly feel that other multi-modal problems will also likely benefit from the proposed attention modules.

2. Related works

Multimodal learning: In RS, the fusion of multiple modalities plays an important role, especially for land-cover classification. Models traditionally based on approaches like Cross-kernel [2], Markov relation [24], morphological [28] and attribute [11] profiles, weighted median filter-based gram Schmidt transform (WMFGS) [37] etc. have shown initial success in exploiting cues from multimodal RS data and provide better classification maps. Later, deep learning methods replaced these ad-hoc approaches with their data-driven feature learning capabilities. In this regard, several works [51, 44, 45, 7] proposed CNN fusion architectures by considering the effects of fusing information at different representation levels: early, middle or late fusion, respectively. In contrast, [12] used both feature-level and decision-level fusion techniques simultaneously to combine the HSI and LiDAR data in Co-CNN. A self-supervised learning-

based HSI-MSI data fusion is proposed in [10]. Similarly, X-ModalNet [19] jointly used self-adversarial, interactive learning, and label propagation modules for cross-modal RS classification. CCR-Net [49], a compact way to fuse heterogeneous RS features for better information exchange.

There are few existing works on RS audiovisual deep learning. DVAN [32] learns the correspondence between the audio and visual modalities in cross-modal retrieval of RS images. The clustering-based *aural atlas* [40] has been built on fusing the audiovisual information. The crowd counting network was designed using joint audiovisual information in [21]. Besides, [16] proposed a self-supervised learning-based approach to understand the key mapping between the RS audiovisual samples and extended it to other transfer learning tasks such as scene classification [20], semantic segmentation [8], cross-modal retrieval [5, 4], etc. In [20], the authors enforced sound-image pairs to transfer the sound event information for RS scene classification. While the existing models are designed for specific pairs of modalities, GAF-Net is generic and can be adapted to any pair of modalities by restructuring the feature extractors.

Attention learning: The usage of attention learning within the CNN frameworks has been proven advantageous in multiple scenarios. Researchers have proposed many easy-to-plugin self-attention modules to highlight the important and non-redundant spatial and spectral feature maps. Generally speaking, there are two variants for the self-attention based models: CNN coupled with self-attention plugins [31, 36, 48], and the vision transformer-based models [9], respectively, and GAF-Net falls under the first category. Squeeze-and-excitation (SE) block [22] provides channel attention by re-calibrating the channel-wise features. A non-local operation-based self-attention module is proposed [46] to capture long-range dependencies in any deep CNN models. Convolutional block attention module (CBAM) [48] and the block attention module (BAM) [36] merge the individually trained channel and spatial attention maps. Residual-based spectral-spatial attention network (RSSAN) [14] for classifying HSI data, where the spectral and spatial attentions help select prominent bands and spatial information, respectively. SSAtt [13] weightily fused the spectral and spatial attention branches. CBAM [48] learns spatial and channel properties from localized transformations, whereas we adopt a non-local approach based on pixel/channel correlation for learning attentive features. GLAM [41] proposed spatial and channel attention based modules to extract the local and global features. We further extend Transformer's [9, 3] spatial attention by incorporating the channel attention modules.

There are studies where cross-attention supports self-attention in multimodal learning such as MCA-Net [27], FusAtNet [33], MBT [35], self-attention based multimodal fusion [54], etc. MCA-Net [27] proposed the optical-SAR-based cross-attention module to generate the joint attention maps. By generating self-attended feature maps and incorporating cross-attention features from LiDAR data in [33], improves the land-cover classification for HSI data. We introduce the novel notion of cross-modal attention learning from single-modal classification networks as opposed to the literature. The existing cross-attention learning model closest to us is [30], which distills the motion attention from a teacher network to a 3D-CNN for human activity network. Clearly, [30] does not concern cross-modal information like GAF-Net; hence, it cannot be directly adopted for the cross-modal fusion task in any context.

3. Proposed methodology

Preliminaries: Let $\mathcal{D} = \{\mathcal{X}_1, \mathcal{X}_2; \mathcal{Y}\}$ be the multimodal dataset, where $\mathcal{X}_1$ and $\mathcal{X}_2$ represent a pair of modalities (such as Audio-Visual, HSI-LiDAR, etc.) and $\mathcal{Y}$ is their respective label space. Further, let ${x^i}_1 \in \mathcal{X}_1$ and ${x^i}_2 \in \mathcal{X}_2$ be the i^{th} input sample point and y^i is its associated label. Under this setup, our goal is to obtain a fused feature representation $z^i = Fe(x_1^i, x_2^i)$, $z^i \in \mathcal{Z}$, for learning an improved classifier: C: $\mathcal{Z} \rightarrow \mathcal{Y}$. In the following, we detail the model architecture for GAF-Net, where both Fe and C are simultaneously learned.

3.1. Model architecture

As illustrated in Figure 2, the GAF-Net architecture consists of two major sub-networks: i) two separate modality-specific classification networks: $\mathcal{T}_1$ and $\mathcal{T}_2$ for $\mathcal{X}_1$ and $\mathcal{X}_2$, and ii) the bi-stream fusion network $\mathcal{S}$ where each of the streams is dedicated to extracting features from a specific input modality. By design, $\mathcal{T}_1$ and $\mathcal{T}_2$ comprise of the deep feature extractors $(Fe^{\mathcal{T}_1}, Fe^{\mathcal{T}_2})$ followed by the classifiers $(C^{\mathcal{T}_1}, C^{\mathcal{T}_2})$, respectively. The main goal of $\mathcal{T}_1$ and $\mathcal{T}_2$ is to learn high-level and discriminative modality-specific features, which can subsequently be utilized to generate the cross-attention masks. On the other hand, the modality-specific feature extractors $Fe_1^{\mathcal{S}}$ and $Fe_2^{\mathcal{S}}$ of $\mathcal{S}$ concatenate the feature-map outputs from each of the self-attended conv. layers with the proposed self-attention block (SAB) and pass them through the 1×1 conv. layer for reducing the dimensions. Note that the feature-map outputs of the intermediate conv. layers are resized via dimension matching block (DM), a global average pooling (GAP) operator that is used to downsample the spatial resolutions of the feature maps to the spatial resolution of the feature maps, which are the output of the final layer of the encoder backbone.

The different conv. layers produce features at different complexities (low, mid, or high-level features), and we feel that considering them together would capture more discriminative aspects from the data. Two separate modality-specific self-attention blocks (SAB) are applied on $Fe_1^{\mathcal{S}}(\mathcal{X}_1)$ and $Fe_2^{\mathcal{S}}(\mathcal{X}_2)$ and the final self-attended

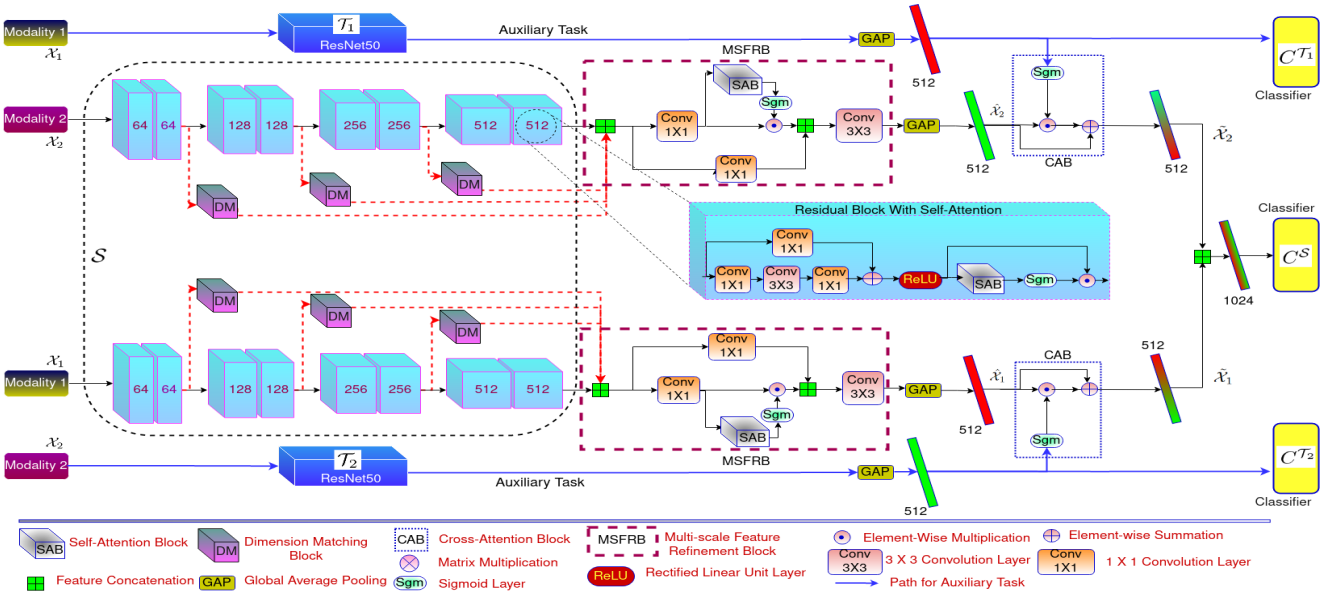

Figure 2. The proposed GAF-Net architecture for multimodal fusion for remote sensing image classification. Here, $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{S}$, represent modality-specific and bi-stream fusion classification networks Incorporated with our proposed spectral-spatial self-attention block (SAB).

features are obtained as $\hat{\mathcal{X}}_1 = GAP(CNN(Fe_1^{\mathcal{S}}(\mathcal{X}_1) \oplus SAB(Fe_1^{\mathcal{S}}(\mathcal{X}_1))))$ and $\hat{\mathcal{X}}_2 = GAP(CNN(Fe_2^{\mathcal{S}}(\mathcal{X}_2) \oplus SAB(Fe_2^{\mathcal{S}}(\mathcal{X}_2))))$, respectively, where $\oplus$ is used to implement the residual connection, and GAP denotes the global average pooling over the depth dimensions. In this way, $\hat{\mathcal{X}}_1$ and $\hat{\mathcal{X}}_2$ are constrained to learn important modality-specific features representations of different complexity; however, there is some latent information encoded in these features which are not expressed naturally. For example, the elevation information from LiDAR data can aid in dense prediction tasks like semantic segmentation. However, a segmentation network by itself overlooks this critical aspect. Under this premise, we aim to refine $\hat{\mathcal{X}}_1$ and $\hat{\mathcal{X}}_2$ considering the cross-modal information. For the same, we generate the cross-attention masks by applying the cross-attention blocks (CAB) on the features obtained from $Fe^{\mathcal{T}_1}(\mathcal{X}_1)$ and $Fe^{\mathcal{T}_2}(\mathcal{X}_2)$, respectively. Henceforth, the final modality specific-features in $\mathcal{S}$ are obtained as:

$$\tilde{\mathcal{X}}_1 = \hat{\mathcal{X}}_1 \oplus (\hat{\mathcal{X}}_1 \odot CAB(Fe^{\mathcal{T}_2(\mathcal{X}_2)}),$$
$$\tilde{\mathcal{X}}_2 = \hat{\mathcal{X}}_2 \oplus (\hat{\mathcal{X}}_2 \odot CAB(Fe^{\mathcal{T}_1(\mathcal{X}_1)}) \quad (1)$$

While the sub-networks of $\mathcal{S}$ till the generation of $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{X}}_2$ define $\mathcal{F}$, the classification network $C^{\mathcal{S}}$ defines f. In the following, we detail the architectures of CAB and SAB, respectively.

3.2. Proposed non-local spectral-spatial self-attention block (SAB)

Our self-attention block uses single-head key-value-based spatial attention (SA) and channel attention (CA) modules. Furthermore, as shown in Figure 3 (d), channel attention comprises global channel attention (GCA) and local channel attention (LCA), which help in extracting important channel attributes for the pixels from two different viewpoints, one using the global channel context while the other exploiting different local channel-wise spatial information effectively. CA and SA modules work on the same input feature maps, and the obtained outputs are summed up element-wise. Similarly, the outputs of both CA modules are element-wise added. We apply the SAB after the individual convolution blocks and on the combined multi-level feature outputs from all the convolution layers separately for each encoder. In a way, SAB performs a multi-level feature refinement (MSFRB) and aggregation, thus highlighting the important feature hierarchy per domain.

Spatial attention (SA): This attention module is designed to learn the insightful spatial features by taking both the short-range and long-range pixel interactions from the input feature maps (Figure 3 (a)). Here, the same input feature maps with dimensions $\mathbb{R}^{C\times H\times W}$ (C, H, W define the channel, height, and width of the feature maps) are provided to the value (V), key (K), and query (Q) tensors and are first fed to the 1×1 conv. layer for dimension reduction. This is to compensate for the multiple heads which process the non-overlapping set of features in the traditional multi-head attention blocks [43]. Precisely, we first down-sample the channels of K and Q by eight times, i.e., $\mathbb{R}^{C/8\times H\times W}$ and then flatten the height and width dimensions to form the tensor of dimensions $\mathbb{R}^{C/8\times HW}$. Subsequently, we create an attention mask of size $\mathbb{R}^{HW\times HW}$ using matrix multiplication ($\otimes$) between K and transposed Q features. We then finally pass the attention mask to the softmax activation layer followed by matrix multiplication cum addition in a residual manner with V to obtain the spatially attended output feature maps as defined $V + V \otimes softmax(K \otimes Q^T)$, with dimensions $\mathbb{R}^{C\times H\times W}$ (Figure 3a). In contrast to [43], we induced the residual connection here to create stability dur-

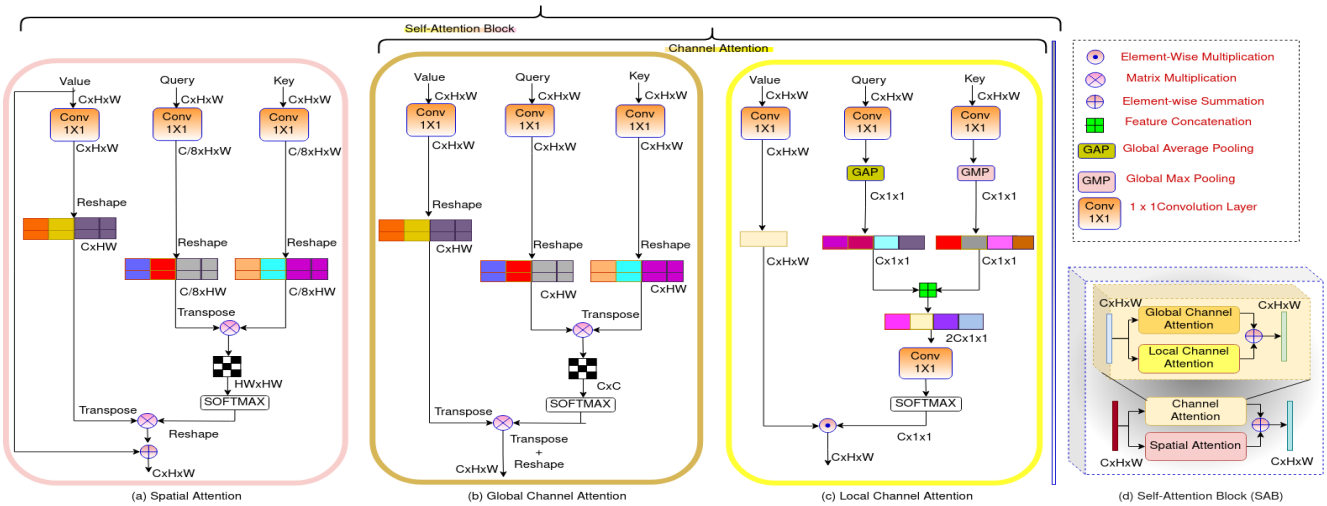

Figure 3. Our proposed self attention block, which consists of (a) spatial-attention (SA), (b) global channel attention (GCA), (c) local channel attention (LCA) modules and (d) implementation of SAB.

ing gradient flow.

Global channel attention (GCA): From Figure 3 (b), it is imperative that this attention module highly resembles the spatial attention block in design except for the residual path and down-sampling of channels of Q and K, which are not considered here. Specifically, we propose to compute the channel-wise attention mask with dimensions $\mathbb{R}^{C\times C}$ by matrix multiplying the 1×1 convolved Q and K feature matrices over the $\mathbb{R}^{C\times H\times W}$ dimensional input feature maps. We pass these attention masks through the softmax layer and then perform matrix multiplication ($\otimes$) with the reshaped V matrix of shape $\mathbb{R}^{C\times HW}$ to obtain the final $\mathbb{R}^{C\times H\times W}$ dimensional attended features. The primary motivation of this channel attention is to assess the cross-correlation between a given channel and all the other channels spectrally attending a given pixel, thus providing a global context for the channel dimensions. To our knowledge, such a paradigm has not been considered as the tradition is to attend/weight the channel dimensions independently. In gist, the attended feature maps are calculated as:

$$\mathcal{R}((\mathcal{R}(conv_{1\times1}(V))\otimes softmax(\mathcal{R}(conv_{1\times1}(Q))\otimes\mathcal{R}(conv_{1\times1}(K))^T))^T) \quad (2)$$

where $\mathcal{R}$ defines the reshape operation applied to tensors.

Local channel attention (LCA): This module locally attends to the channel dimensions instead of global channel attention. As per Figure 3 (c), the V, Q, and K feature vectors are passed through the 1×1 convolution layer as spatial attention and global channel attention, but Q and K feature vectors are additionally passed with global average pool (GAP) and global max pool (GMP) layers to get the dimensions of $\mathbb{R}^{C\times1\times1}$, respectively, to highlight the spatial contexts. This way, the high and low-frequency spatial information over each channel dimension is encoded. We concatenate Q and K and compress the fused features using 1×1 conv. layer. Finally, we pass the compressed features with dimensions $\mathbb{R}^{C\times1\times1}$ through the softmax activation layer and use element-wise multiplication ($\odot$) with the V feature vector to obtain the attended feature maps with dimensions $\mathbb{R}^{C\times H\times W}$ as follows,

$$conv_{1\times1}(V)\odot softmax(conv_{1\times1}(CONCAT(GAP(conv_{1\times1}(Q)),GMP(conv_{1\times1}(K))))) \quad (3)$$

3.3. Across-modality cross-attention block (CAB)

The discriminative cross-modal information further revives these self-attended features to highlight some informative hidden feature properties. Note that CAB works on the vector-valued intermediate high-level semantic representations $\hat{\mathcal{X}}_1$ and $\hat{\mathcal{X}}_2$, respectively. First, $Fe^{\mathcal{T}_1}(\mathcal{X}_1)$ / $Fe^{\mathcal{T}_2}(\mathcal{X}_2)$ are designed to match the length of $\hat{\mathcal{X}}_1$ / $\hat{\mathcal{X}}_2$. Henceforth, we pass $Fe^{\mathcal{T}_1}(\mathcal{X}_1)$ through GAP and sigmoid layers to generate the attention masks, which are then element-wise multiplied ($\odot$) with $\hat{\mathcal{X}}_2$ together with adding a residual connection. A similar process is followed for the other modality, and we obtain $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{X}}_2$ (Eq. 1).

3.4. Objective function for training GAF-Net

This section defines the loss functions to train our proposed GAF-Net in an end-to-end manner. $\mathcal{T}_1$ and $\mathcal{T}_2$ are trained only for $(\mathcal{X}_1,\mathcal{Y})$ and $(\mathcal{X}_2,\mathcal{Y})$ using two cross entropy (CE) losses $\mathcal{L}_1$ and $\mathcal{L}_2$ respectively. On the other hand, we concatenate $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{X}}_2$ to obtain $\mathcal{Z}$ and define another CE loss $\mathcal{L}_3$ on $C^{\mathcal{S}}$. We would further ensure non-redundancy between $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{X}}_2$; we propose a non-redundancy regularizer $\mathcal{L}_{NRR}$ which tends to minimize the cross-correlation between the l_2 normalized representations of $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{X}}_2$, $\tilde{\mathcal{X}}_1'$ and $\tilde{\mathcal{X}}_2'$, as follows,

$$\mathcal{L}_{NRR}=||\tilde{\mathcal{X}}_1'^T\tilde{\mathcal{X}}_2'-\mathbb{I}||_2 \quad (4)$$

where $\mathbb{I}$ denotes the identity matrix. This loss constraint the cross-correlation terms to take the value of zero, thus making both the modality-specific features look into non-overlapping aspects regarding the input data. The overall

Table 1. Comparison of our proposed GAF-Net with SOTA methods on Houston 2013 HSI-LiDAR and HSI-MSI, Berlin HSI-SAR, and Augsburg HSI-SAR hyperspectral datasets. $^{\$}$ represent multi inputs with End-Net [17]. * and ** are the performance of modality-specific networks $\mathcal{T}_1$ and $\mathcal{T}_2$ on HS modality and other modalities, respectively. $\mathcal{T}_1 + \mathcal{T}_2$ denotes fusion of $Fe^{\mathcal{T}_1}$ and $Fe^{\mathcal{T}_2}$. $\pm$ represents the standard deviation. The results in $^{\#}$ uses Transformer as the feature extractor, whereas we used CNN blocks for feature extraction, which is not a fair comparison. We highlighted the best results in **bold**.

Methods	Houston2013 HSI-LiDAR			Houston2013 HSI-MSI			Berlin HSI-SAR			Augsburg HSI-SAR		
	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ
Hyper-Embedder [50]	82.64±0.45	83.15	0.8070	82.77±0.30	83.81	0.8134	59.98±0.38	60.12	0.4641	81.03±0.21	52.56	0.7243
Two-Branch CNN [51]	87.98±0.29	90.11	0.8698	86.56±0.23	89.21	0.8546	63.73±0.20	62.34	0.4904	83.57±0.33	62.10	0.7723
End-Net $^{\$}$[17]	88.52±0.24	89.85	0.8759	87.65±0.40	88.29	0.8610	64.01±0.32	61.88	0.5001	84.11±0.15	62.78	0.7758
Co-CNN [12]	88.96±0.41	89.21	0.8766	85.44±0.33	84.10	0.8237	64.08±0.26	62.83	0.5925	87.76±0.52	62.71	0.8040
CCR-Net [49]	89.66±0.27	91.53	0.8877	88.15±0.19	89.82	0.8719	69.85±0.46	66.99	0.5716	86.32±0.28	64.47	0.8003
FusAtNet [33]	89.98±0.34	94.65	0.8913	86.17±0.51	86.39	0.8408	63.45±0.29	63.19	0.5088	84.42±0.30	62.66	0.7782
S2FL [18]	-	-	-	85.07±0.23	86.11	0.8378	62.23±0.19	62.48	0.4877	83.36±0.22	61.38	0.7626
AsyFFNet [26]	-	-	-	-	-	-	70.51±0.14	70.31	0.5824	89.14±0.27	69.16	0.8452
MFT $^{\#}$[39]	89.80 ±0.53	91.51	0.8893	89.15±0.96	90.56	0.8822	-	-	-	90.49±0.20	60.36	0.8626
$\mathcal{T}_1^{\star}$	86.02±0.32	88.56	0.8481	86.02±0.24	88.56	0.8481	68.11±0.47	54.61	0.5951	84.08±0.21	58.21	0.7759
$\mathcal{T}_2^{\star\star}$	67.27±0.19	70.66	0.6693	75.00±0.40	78.84	0.7298	64.66±0.23	36.30	0.3694	84.07±0.33	50.46	0.7700
$\mathcal{T}_1$+$\mathcal{T}_2$	87.09±0.23	89.15	0.8511	87.99±0.34	89.43	0.8564	69.32±0.19	58.77	0.6003	85.75±0.25	60.89	0.7788
GAF-Net	**91.39±0.21**	**94.92**	**0.9018**	**90.64±0.17**	**93.30**	**0.8938**	**78.57±0.23**	**70.92**	**0.6761**	**90.80±0.12**	**70.10**	**0.8683**

Table 2. Comparison of our proposed GAF-Net with SOTA methods on the ADVANCE dataset. $^{\#}$ TL represent Triplet Loss from [16]. $\pm$ represents the standard deviation. We highlighted the best results in **bold**.

Method	Audio Baseline [20]	Visual Baseline [20]	Audio-visual Baseline [20]	Batch TL $^{\#}$ [16]	Audio $\mathcal{T}_1$	Visual $\mathcal{T}_2$	$\mathcal{T}_1$+$\mathcal{T}_2$	**GAF-Net**
Precision	30.46±0.23	74.05±0.31	75.25±0.27	89.59±0.19	73.28±0.55	89.48±0.43	89.90±0.20	**93.37±0.11**
Recall	32.99±0.46	72.79 ±0.25	74.79±0.11	89.52±0.18	73.50±0.41	89.34±0.25	90.21±0.37	**93.23±0.21**
F1	28.99±0.39	72.85±0.27	74.58±0.40	89.50±0.33	73.38±0.21	89.40±0.45	90.05±0.30	**93.31± 0.17**

multi-task loss function is defined in Eq. 5.

$$\mathcal{L}_{Total} = \sum_{i=1}^{3} \mathcal{L}_i + \mathcal{L}_{NRR} \tag{5}$$

4. Experimental protocols

Houston 2013 HSI-LiDAR: The National Centre for Airborne Laser Mapping (NCALM) introduced this data in the GRSS Data Fusion Contest 2013, covering Houston University and its nearby surroundings. It has 144 spectral bands ranging from 0.38 μm to 1.05 μm. Each channel consists of a 349×1905 pixel raster map with a spatial resolution of 2.5 m and one LiDAR band with the same raster size as the HSI bands. Training and test sets of 2832 and 12197 pixels are provided for this dataset [33].

Augsburg HSI-SAR: The original data [18] is composed of three different modalities; HSI, SAR, and digital surface model (DSM), out of which a pair is considered at a time to define three fusion tasks: HSI-SAR, HSI-DSM, and SAR-DSM, respectively. We concentrate only on HSI-SAR fusion and compare it with different baselines for 7-classes. The scenes have a spatial resolution of 30 m in dimension 332×485 with 180 spectral bands between 0.4 μm to 2.5 μm, DSM image with one band, and four features from dual-Pol SAR image. It contains 78294 samples, of which 761 and 77533 are used for training and testing.

Berlin HSI-SAR: The dataset [18] consists of HSI and SAR scenes with a resolution of 1723×476 pixels from eight land-cover classes of the urban and rural areas surrounding the city of Berlin. The total sample count of 464671 is divided into 2820 for training and 461851 for testing. [18] mentions image pre-preprocessing for HS and SAR.

Houston 2013 (HSI-MSI): From the original HSI and MSI images, [18] created the multimodal data. It has the same spectral and spatial resolutions as the Houston 2013 HSI-LiDAR dataset and the same classes and samples.

ADVANCE: To better assess the generalization of our proposed GAF-Net, we consider a different set of multimodal data apart from HSI images. The dataset presented in [20] consists of 5075 pairs of audio-visual samples, from which 4056 samples are used for training and the remaining for testing, followed by a 5-fold cross-validation.

4.1. Model architecture and training protocols

In our GAF-Net architecture, the deep feature extractors $Fe^{\mathcal{T}_1}$ and $Fe^{\mathcal{T}_2}$ consist of ResNet-50 [15] architecture and provide the linear feature embeddings of dimension 512. Whereas, the modality-specific feature extractors $Fe_1^{\mathcal{S}}$ and $Fe_2^{\mathcal{S}}$ of $\mathcal{S}$ are made up of four pairs of residual-based conv. blocks that compute feature maps with depths of 64, 128, 256, and 512, respectively. Furthermore, within each of these conv. blocks, the outputs are self-attended by SAB modules, and the application of a residual connection produces the output feature maps. To ensure stable training, we employ ReLU non-linearity and Batch-normalization after each conv. block of $Fe_1^{\mathcal{S}}$ and $Fe_2^{\mathcal{S}}$. Finally, the classifiers $C^{\mathcal{T}_1}, C^{\mathcal{T}_2}$, and $C^{\mathcal{S}}$ take linear feature embeddings of $512, 512$, and 1024, respectively.

Here, we mention the training strategies with standard settings similar to [33, 12, 18] on the multimodal HSI data and followed to use the cubical patches of size 17×17 around each pixel with the ground-truth label for all the HSI, MSI, and SAR images. Subsequently, PCA [38] is used to reduce the channel dimensionality of the HSI data to the dimension of 30, and remove any redundant band information. Training is performed using ADAM optimizer [25] with an initial learning rate of 10^{-2}, and for every 40 epoch, the scheduler decreases the learning rate by a factor of 10^{-1} and a total of 200 training epochs are performed given a mini-batch size of 16. For the audio modality in the ADVANCE dataset, following [20, 16], the spectrograms are generated with dimensions 400×64, whereas the images from the visual modality are resized to 256×256. In order to compensate for data and class imbalance in ADVANCE, we use augmentation techniques such as random

Table 3. Ablation analysis of our proposed GAF-Net on Houston 2013 HSI-LiDAR and HSI-MSI, Berlin HSI-SAR, and Augsburg HSI-SAR hyperspectral datasets. **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, global channel attention, and local channel attention, respectively. We highlighted the best results in **bold**.

Methods					Houston2013 HSI-LiDAR			Houston2013 HSI-MSI			Berlin HSI-SAR			Augsburg HSI-SAR		
A: Baselines					OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ
	Layer-wise SAB	CAB	$\mathcal{L}_{NRR}$	MSFRB												
A1:	✗	✗	✗	✗	88.32	90.08	0.8731	84.51	87.21	0.8381	71.78	64.79	0.6421	82.64	61.80	0.7779
A2:	✗	✗	✓	✗	89.10	90.89	0.8791	85.11	87.97	0.8414	72.38	64.97	0.6432	82.99	62.20	0.7808
A3:	✓	✗	✗	✗	88.97	90.77	0.8779	84.91	87.56	0.8388	71.95	64.83	0.6422	82.78	61.99	0.7801
A4:	✓	✗	✓	✗	89.31	91.37	0.8811	85.35	88.85	0.8484	72.94	65.37	0.6482	83.03	62.59	0.7851
A5:	✗	✓	✗	✗	89.06	90.99	0.8800	84.95	88.37	0.8410	71.88	64.73	0.6399	83.17	62.22	0.7832
A6:	✗	✗	✗	✓	89.11	91.25	0.8826	84.95	87.86	0.8405	72.32	64.97	0.6455	83.17	62.39	0.7866
A7:	✗	✓	✓	✗	89.98	91.93	0.8879	89.02	90.67	0.8821	72.31	65.01	0.6449	85.50	64.80	0.8033
A8:	✗	✓	✓	✓	90.41	93.00	0.8918	89.08	90.45	0.8819	73.45	65.33	0.6574	85.95	65.11	0.8098
A9:	✓	✓	✗	✓	90.55	93.18	0.8937	89.13	90.66	0.8829	73.39	65.02	0.6512	85.59	64.99	0.8041
A10:	✓	✓	✓	✗	90.35	92.55	0.8902	89.74	91.80	0.8887	75.70	67.61	0.6604	86.21	65.17	0.8112
B: Ablation on SAB																
B1:	SA				90.11	93.02	0.8919	89.23	91.67	0.8823	76.88	67.81	0.6641	87.44	67.50	0.8323
B2:	GCA				89.99	92.91	0.8895	89.05	91.04	0.8789	76.75	67.59	0.6606	87.51	67.70	0.8341
B3:	LCA				89.37	91.45	0.8831	88.73	90.85	0.8759	75.98	66.94	0.6591	86.99	67.01	0.8291
B4:	SA + GCA				90.66	93.77	0.8987	89.91	92.89	0.8878	77.44	68.17	0.6698	88.25	68.43	0.8410
B5:	SA + LCA				90.32	93.11	0.8965	89.80	92.55	0.8833	77.03	67.85	0.6666	87.98	68.22	0.8399
B6:	SA + GCA + LCA †				90.69	93.80	0.8975	89.73	92.60	0.8841	77.32	68.16	0.6671	88.33	68.28	0.8429
GAF-Net					**91.39**	**94.92**	**0.9018**	**90.64**	**93.30**	**0.8938**	**78.57**	**70.92**	**0.6761**	**90.80**	**70.10**	**0.8683**

Table 4. Ablation analysis on CAB for Houston 2013 HSI-LiDAR and HSI-MSI, Berlin HSI-SAR, and Augsburg HSI-SAR hyperspectral datasets. IL in * and FL in ** denote Intermediate Layer and Final Layer, respectively. We highlighted the best results in **bold**.

Methods	Houston2013 HSI-LiDAR			Houston2013 HSI-MSI			Berlin HSI-SAR			Augsburg HSI-SAR		
CAB Ablation	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ
No CAB	89.31	91.37	0.8811	85.35	88.85	0.8484	72.94	65.37	0.6482	83.03	62.59	0.7851
CAB from IL of $\mathcal{T}_1$ and $\mathcal{T}_2{}^{*}$	89.55	91.65	0.8844	87.79	89.10	0.8611	73.99	66.89	0.6579	84.81	64.44	0.8008
CAB from FL of $\mathcal{S}$ **	89.11	90.97	0.8798	86.18	89.46	0.8511	72.33	64.76	0.6419	83.12	62.48	0.7865
GAF-Net	**91.39**	**94.92**	**0.9018**	**90.64**	**93.30**	**0.8938**	**78.57**	**70.92**	**0.6761**	**90.80**	**70.10**	**0.8683**

90° rotation, random hue-saturation value shifting, and random horizontal and vertical flips with random shifting with $+90^{\circ}$ to -90° random rotation, with a probability of 0.5 for each augmentation. We train our network with a batch size of 16 in a scheduled way for 300 epochs, i.e., the Adam [25] optimizer is set to a learning rate of 10^{-3} for the first 100 epochs, and the learning rate is sliced by 10 times for each subsequent 100 epochs.

We adopt overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ) for reporting the mean accuracies over three runs for all the datasets concerning only the visual modalities. For ADVANCE, we use f1-score (F1), precision, and recall as the evaluation metrics, following the relevant literature.

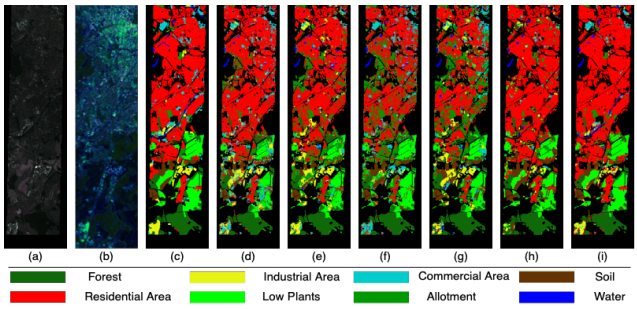

Figure 4. Classification maps for Berlin HSI-SAR dataset. (a) HSI's true colour composite (RGB Bands - 43, 22, 36), (b) SAR's true color composite (RGB Bands- 2, 1, 4), and (c) Ground-truth. From (d) to (i) represent classification maps for different methods, i.e., (d)End-Net [17], (e) Co-CNN [12], (f) CCR-Net [49], (g) FusAtNet [33], (h) $\mathcal{T}_1 + \mathcal{T}_2$, (i) GAF-Net

4.2. Comparison to the literature

We evaluate the performance of GAF-Net with benchmark methods on multimodal learning from the RS community dealing with visual, DSM, and audio data, as illustrated in Tables 1 and 2. Specifically, we consider HyperEmbedder [50], Two-Branch CNN [51], multi-input End-Net [17], Co-CNN [12], CCR-Net [49], FusAtNet [33], S2FL [18], and AsyFFNet [26] where an HSI is involved. It can be observed that GAF-Net consistently outperforms the referred SOTA methods. On Houston2013 HSI-LiDAR and HSI-MSI, Berlin HSI-SAR, and Augsburg HSI-SAR, our GAF-Net is found to enhance the OA by at least 1.9%, 2.7%, 11%, and 1.7%, AA by 0.2%, 3.7%, 5.5%, and 7.5%, and 1.7%, and κ by at least 1.2%, 3.7%, 12%, and 5%, respectively, from the closest literature. Amongst these methods, note that FusAtNet [33] integrates both self-attention and cross-attention modules using the conventional CNN-based networks; however, the proposed SAB and CAB architectures seem to be helpful in this regard. Finally, we compare the performance of the fusion sub-network $\mathcal{S}$ with the outputs of $\mathcal{T}_1$ and $\mathcal{T}_2$ from two perspectives: i) when they are trained using two different classifiers, and ii) they are trained with a single classifier ($\mathcal{T}_1 + \mathcal{T}_2$). The sub-network $\mathcal{S}$, although it is shallow as compared to $\mathcal{T}_1$ / $\mathcal{T}_2$, beats all these baselines convincingly.

On the other hand, we record 4% increment from the Batch TL [16] for the audio-visual ADVANCE dataset. Here, we also compare our proposed architecture with $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_1 + \mathcal{T}_2$ and our GAF-Net effectively improved the precision by a margin of 4%. In Figure 4, we generate the classification maps for the Berlin HSI-SAR dataset.

4.3. Architecture ablation and loss functions

We analyze the effects of each component of GAF-Net in Table 3. We first consider the base model without CAB, SAB, and $\mathcal{L}_{NRR}$, and then incrementally consider all the model components and observe that applying these com-

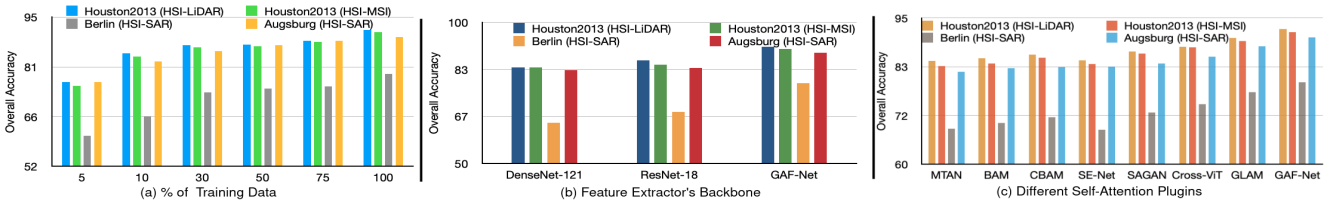

Figure 5. Analysis (a) % of used training data, (b) feature extractor backbones for $\mathcal{S}$, and (c) different self-attention plugins in GAF-Net on the HSI datasets.

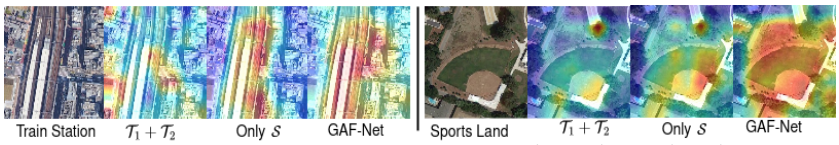

Figure 6. The generated class activation map from the fusion modality-specific network $\mathcal{T}_1$ and $\mathcal{T}_2$, i.e., $\mathcal{T}_1+\mathcal{T}_2$, sub-network $\mathcal{S}$ (only $\mathcal{S}$ represents S with SAB no CAB) and GAF-Net on the ADVANCE dataset.

ponents progressively improves the performance. An enhancement of at least $4-5\%$ can be observed in the OA values from the naive baseline to the full GAF-Net. Furthermore, we ablate the spatial attention (SA) and channel attention blocks (GCA and LCA) of our SAB module. Similar to the previous case, we consider their individual effects followed by combined effects and confirm that all of the SAB modules are important in GAF-Net.

4.4. Critical analysis

Analysis of CAB: In order to 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 2% (Houston2013 HSI-LiDAR), 3.1% (Houston2013 HSI-MSI), 5.8% (Berlin HSI-SAR) and 5% (Augsburg HSI-SAR) in OA values.

Sensitivity to the amount of training samples: In Figure 5 (a), we varied the training size for all multimodal hyperspectral datasets from 5% to 100% of the available training samples. It can be observed that the performance is not significantly degraded even in the low-data regimes.

Analysis of $Fe_1^{\mathcal{S}}$ and $Fe_2^{\mathcal{S}}$: For $\mathcal{S}$, we ablate the feature extractors' backbones and compare $Fe_1^{\mathcal{S}}$ and $Fe_2^{\mathcal{S}}$ against DenseNet-121 [23] and ResNet-18 [15] in Figure 5 (b). The motive is to see the effects of combining features from all the layers as done in GAF-Net and to see whether the shallow $Fe_1^{\mathcal{S}}$ and $Fe_2^{\mathcal{S}}$ of $\mathcal{S}$ can provide comparable performance measures against deeper and more complex backbones with the applications of CAB and SAB. Numerically, our GAF-Net outperforms the referred backbone architectures minimum by 1.5%, 1.8%, 8%, and 4.3% in the OA values for Houston2013 HSI-LiDAR, Houston2013 HSI-MSI, Berlin HSI-SAR, and Augsburg HSI-SAR datasets, respectively. [1]

[1]Supplementary paper contains ablations on the ADVANCE dataset.

Comparison on the attention modules: Finally, Figure 5 (c) shows the performance comparisons when different benchmark attention plugins are used in place of our proposed self-attention block (SAB): CBAM [48], BAM [36], SE-Net [22], MTAN [31], GLAM [41], and SAGAN [52], respectively. Many of them use only spatial attention, while others use spectral-spatial attention learning. We also consider the Cross-ViT [3] attention given the superior performance in different vision tasks. The attention in GAF-Net is found to outperform all of these attention plugins convincingly by $3-5\%$. In supplementary, we provide the effect of agnostic nature of our proposed attention modules with backbones like DenseNet-121 [23] and ResNet-18 [15].

Class Activation Map (CAM) Visualization: In Figure 6, we present the CAM [53] on scenes from the ADVANCE dataset. The CAM clearly suggests that the full GAF-Net can explain the classes intuitively by focusing on the relevant image parts.

5. Conclusions

This paper presents a novel multimodal fusion architecture (GAF-Net) for RS data that uses a novel attention-distillation-based cross-attention between the cross-modal features while novel single-head spectral-spatial self-attention-based feature learning for the individual modalities. The proposed spectral attention considers both the local and global channel contexts, thus extracting better channel features, which is essential for RS data. Though shallower than the individual modality-specific classifiers, our fusion network is found to beat them by a large margin. We show extensive experiments on five multimodal remote sensing benchmark datasets consisting of HSI, MSI, SAR, LiDAR, and audio data and confirm the efficacy of GAF-Net all through. The future scope may consider extending GAF-Net to support more than two modalities. One possibility could be to follow a sequential approach where two modalities are fused first, and the obtained representations are fused with a new modality and so on.

References

[1] Charlotte A. Bishop, Jian Guo Liu, and Philippa J. Mason. Hyperspectral remote sensing for mineral exploration in pulang, yunnan province, china. *International Journal of Remote Sensing*, 32(9):2409–2426, 2011.

[2] Gustavo Camps-Valls, Luis Gomez-Chova, Jordi Munoz-Mari, JosÉ Luis Rojo-Alvarez, and Manel Martinez-Ramon. Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6):1822–1835, 2008.

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

[4] Yaxiong Chen and Xiaoqiang Lu. A deep hashing technique for remote sensing image-sound retrieval. *Remote Sensing*, 12(1), 2020.

[5] Yaxiong Chen, Xiaoqiang Lu, and Shuai Wang. Deep cross-modal image–voice retrieval in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10):7049–7061, 2020.

[6] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. Multi-head attention: Collaborate instead of concatenate. *CoRR*, abs/2006.16362, 2020.

[7] Camille Couprie, Clement Farabet, Laurent Najman, and Yann Lecun. Indoor semantic segmentation using depth information. 01 2013.

[8] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *CoRR*, abs/1805.06561, 2018.

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

[10] Jianhao Gao, Jie Li, and Menghui Jiang. Hyperspectral and multispectral image fusion by deep neural network in a self-supervised manner. *Remote Sensing*, 13(16), 2021.

[11] Pedram Ghamisi, Jon Atli Benediktsson, and Stuart Phinn. Land-cover classification using both hyperspectral and lidar data. *International Journal of Image and Data Fusion*, 6(3):189–215, 2015.

[12] Renlong Hang, Zhu Li, Pedram Ghamisi, Danfeng Hong, Guiyu Xia, and Qingshan Liu. Classification of hyperspectral and lidar data using coupled cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7):4939–4950, 2020.

[13] Renlong Hang, Zhu Li, Qingshan Liu, Pedram Ghamisi, and Shuvra S. Bhattacharyya. Hyperspectral image classification with attention-aided cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3):2281–2293, 2021.

[14] Juan Mario Haut, Mercedes E. Paoletti, Javier Plaza, Antonio Plaza, and Jun Li. Visual attention-driven hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10):8065–8080, 2019.

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

[16] Konrad Heidler, Lichao Mou, Di Hu, Pu Jin, Guangyao Li, Chuang Gan, Ji-Rong Wen, and Xiao Xiang Zhu. Self-supervised audiovisual representation learning for remote sensing data. *CoRR*, abs/2108.00688, 2021.

[17] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4340–4354, 2020.

[18] Danfeng Hong, Jingliang Hu, Jing Yao, Jocelyn Chanussot, and Xiao Xiang Zhu. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:68–80, 2021.

[19] Danfeng Hong, Naoto Yokoya, Gui-Song Xia, Jocelyn Chanussot, and Xiao Xiang Zhu. X-modalnet: A semi-supervised deep cross-modal network for classification of remote sensing data. *CoRR*, abs/2006.13806, 2020.

[20] Di Hu, Xuhong Li, Lichao Mou, Pu Jin, Dong Chen, Liping Jing, Xiaoxiang Zhu, and Dejing Dou. Cross-task transfer for multimodal aerial scene recognition. *CoRR*, abs/2005.08449, 2020.

[21] Di Hu, Lichao Mou, Qingzhong Wang, Junyu Gao, Yuansheng Hua, Dejing Dou, and Xiao Xiang Zhu. Ambient sound helps: Audiovisual crowd counting in extreme conditions. *CoRR*, abs/2005.07097, 2020.

[22] 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.

[23] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[24] Christian Häne, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. Joint 3d scene reconstruction and class segmentation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 97–104, 2013.

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[26] Wei Li, Yunhao Gao, Mengmeng Zhang, Ran Tao, and Qian Du. Asymmetric feature fusion network for hyperspectral and sar image classification. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022.

[27] Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106:102638, 2022.

[28] Wenzhi Liao, Aleksandra Pižurica, Rik Bellens, Sidharta Gautama, and Wilfried Philips. Generalized graph-based fusion of hyperspectral and lidar data using morphological features. *IEEE Geoscience and Remote Sensing Letters*, 12(3):552–556, 2014.

[29] Liyuan Liu, Jialu Liu, and Jiawei Han. Multi-head or single-head? an empirical comparison for transformer training. *CoRR*, abs/2106.09650, 2021.

[30] Miao Liu, Xin Chen, Yun Zhang, Yin Li, and James M Rehg. Attention distillation for learning video representations. *arXiv preprint arXiv:1904.03249*, 2019.

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

[32] Guo Mao, Yuan Yuan, and Lu Xiaoqiang. Deep cross-modal retrieval for remote sensing image and audio. In *2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*, pages 1–7, 2018.

[33] Satyam Mohla, Shivam Pande, Biplab Banerjee, and Subhasis Chaudhuri. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 416–425, 2020.

[34] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12416–12425, 2019.

[35] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *CoRR*, abs/2107.00135, 2021.

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

[37] Yinghui Quan, Yingping Tong, Wei Feng, Gabriel Dauphin, Wenjiang Huang, and Mengdao Xing. A novel image fusion method of multi-spectral and sar images for land cover classification. *Remote Sensing*, 12(22), 2020.

[38] Craig Rodarmel and Jie Shan. Principal component analysis for hyperspectral image classification. *Surv Land inf Syst*, 62, 01 2002.

[39] Swalpa Kumar Roy, Ankur Deria, Danfeng Hong, Behnood Rasti, Antonio Plazza, and Jocelyn Chanussot. Multimodal fusion transformer for remote sensing image classification. *arxiv.org/abs/2203.16952*, 2022.

[40] Tawfiq Salem, Menghua Zhai, Scott Workman, and Nathan Jacobs. A multimodal approach to mapping soundscapes. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 3477–3480, 2018.

[41] Chull Hwan Song, Hye Joo Han, and Yannis Avrithis. All the attention you need: Global-local, spatial-channel attention for image retrieval. *CoRR*, abs/2107.08000, 2021.

[42] Freek D Van der Meer, Harald MA Van der Werff, Frank JA Van Ruitenbeek, Chris A Hecker, Wim H Bakker, Marleen F Noomen, Mark Van Der Meijde, E John M Carranza, J Boudewijn De Smeth, and Tsehaie Woldai. Multi-and hyperspectral geologic remote sensing: A review. *International Journal of Applied Earth Observation and Geoinformation*, 14(1):112–128, 2012.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[44] Anran Wang, Jiwen Lu, Jianfei Cai, Tat-Jen Cham, and Gang Wang. Large-margin multi-modal deep learning for rgb-d object recognition. *Trans. Multi.*, 17(11):1887–1898, nov 2015.

[45] Wei Wang, Xiaoyan Yang, Beng Chin Ooi, D. Zhang, and Yueting Zhuang. Effective deep learning-based multi-modal retrieval. *The VLDB Journal*, 25:79–101, 2015.

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[47] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10938–10947, 2020.

[48] 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.

[49] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. Convolutional neural networks for multimodal remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–10, 2022.

[50] Xin Wu, Danfeng Hong, and Di Zhao. Hyper-embedder: Learning a deep embedder for self-supervised hyperspectral dimensionality reduction. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[51] Xiaodong Xu, Wei Li, Qiong Ran, Qian Du, Lianru Gao, and Bing Zhang. Multisource remote sensing data classification based on convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):937–949, 2018.

[52] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[53] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.

[54] Hu Zhu, Ze Wang, Yu Shi, Yingying Hua, Guoxia Xu, and Lizhen Deng. Multimodal fusion method based on self-attention mechanism. *Wireless Communications and Mobile Computing*, 2020, 2020.