Neurocomputing 635 (2025) 129855

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

GSNet: A new small object attention based deep classifier for presence of gun in complex scenes

Rajib Debnath [a,b], Kakali Das [a,b], Mrinal Kanti Bhowmik [a,*]

[a] Department of Computer Science and Engineering, Tripura University (A Central University), Suryamaninagar, Tripura 799022, India
[b] Department of Computer Science and Engineering, GITAM (Deemed to be University), Hyderabad Campus, Hyderabad, Telangana, 502329, India

ARTICLE INFO

Communicated by Q. Wang

Keywords:
Weapon
Scene classification
Attention module
Dense network
TUVD-CSA

ABSTRACT

The motivation for focusing on weapon-based scene classification stems from the critical need to enhance public safety by enabling automated systems to quickly and accurately detect firearms in various environments. In contrast to common surveillance scenario classification based on intruders, weapon-based scenario classification often involves small weapons distributed throughout the scene or image. This requires more discriminative features and local semantics for effective classification. However, when deep convolutional neural networks (CNNs) are applied to scene classification, the loss of low- and mid-level features cannot be avoided. Furthermore, most existing networks tend to emphasize the global semantics of images. The low inter-class variability and high intra-class variability present specific challenges in weapon-based scene classification. To address these challenges, we propose a small object attention-based architecture in this work, with DenseNet serving as the backbone of our classification model. We modified the original DenseNet architecture to obtain more structured features. Additionally, we introduce a Small Object Attention (SAN) module after each dense block and an enhancement layer after each transition layer. Furthermore, we propose an enhanced classification layer in place of the traditional softmax layer, which helps retain relevant semantic features during classification. Consequently, the proposed classification model processes small patches of the image, preserving the relevant features of weapons. Experiments on six widely used benchmark datasets for weapon-based scenes demonstrate that our GSNet outperforms state-of-the-art methods by a significant margin while utilizing considerably fewer parameters. On average, the DenseNet model achieved an accuracy of 94.2%, whereas the proposed network attained an average accuracy of 98% across the six datasets.

1. Introduction

The research area of automatic scene classification based on firearms can be categorized as a specific case of the general small object detection problem and is an emerging field with recent advances. The motivation for this research stems from the increasing need for enhanced public safety and security, especially given the rise in firearm-related incidents globally. According to a 2023 report by the Small Arms Survey, there are over 1 billion firearms in circulation worldwide, with civilian-held firearms accounting for 85% of the total [1]. Moreover, firearm-related violence has seen a consistent rise in many countries, with mass shootings and gun-related crimes becoming more frequent in public spaces, such as schools, airports, and government buildings.

Given this alarming trend, the importance of advanced surveillance and security systems cannot be overstated. The priority of "surveillance and security" in the real world is well known, and traditional methods of human monitoring are often insufficient to ensure timely and accurate detection. Surveillance systems embedded with automatic scene classification for firearms offer a critical solution. Such systems can be applied in numerous high-risk areas, including automatic alarm systems for sensitive locations, border security, airport security, the protection of government offices, and safeguarding crowded public spaces such as rallies, political events, and concerts. With the ability to detect small objects like firearms in real-time, these systems not only enhance the responsiveness of security measures but also significantly reduce the reliance on human operators, thus minimizing errors and response times in crisis situations.

Possible challenges in automatic small object detection especially gun detection are as follows:

- **Acentric Distribution of Gun in the given scene:** Compared to the common object detection, in gun detection the distribution is

* Corresponding author.
E-mail addresses: rajibdebnath.cse@gmail.com (R. Debnath), kakalids54@gmail.com (K. Das), mrinalkantibhowmik@tripurauniv.ac.in (M.K. Bhowmik).

https://doi.org/10.1016/j.neucom.2025.129855
Received 20 May 2024; Received in revised form 8 February 2025; Accepted 25 February 2025
Available online 6 March 2025

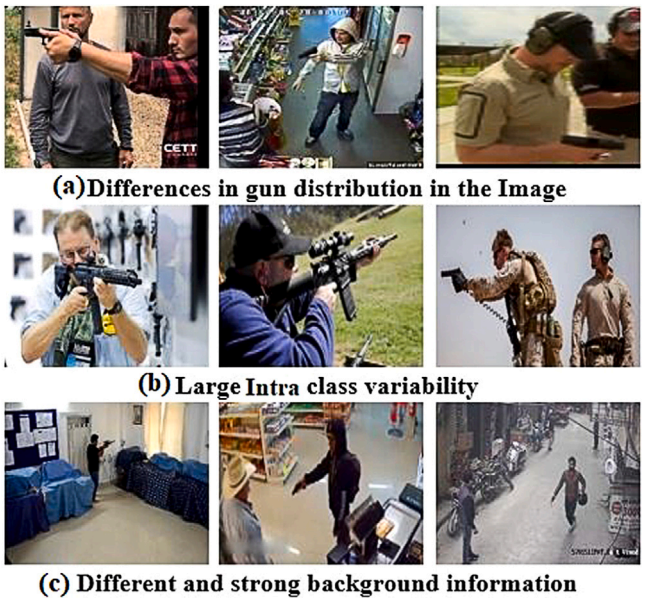

Fig. 1. Different distribution of weapon in input surveillance frame.

not acentric. Gun is a very small object and it can be found in any possible position, in any orientation [2]. Unlike, in other object detection problem where key objects are mostly accommodate in the center of the scene. In Gun detection, gun may be present at corners of the images placed in any table etc. Fig. 1(a) can explain the mentioned challenge using few examples and Fig. 2 showing that in natural images key objects are usually at the center whereas guns can be distributed anywhere in cctv images. Statistics on Imagenet Dataset [3] and in [4] reflect that in the images objects are have distinct differences in the image compared to the guns of olmos et al. Dataset. This characteristics make difficult to understand given scene in terms of guns and robust approaches to this challenge may be qualified.

- **Large Intra class variability:** Classification problem is based on the intra and inter class variability. Problem with high inter class variability is comparably easy to be handled by classification approaches. In gun based classification problem, the inter class variability is low and intra class variability is high [4] as shown in Fig. 1(b). Thus, more discriminative features required for gun based classification.
- **Strong background:** The classification problem can be defined by the problem of determination of scene level based on the key object. But usually scenes have many other objects, local regions irrelevant to the scene level, which is termed as background. Due to different lighting condition, viewing angle, background information can become strong and increase difficulty in detection of the key objects. If the key object is gun than the problem may become worse as there is possibility of large number of false positives [5]. Fig. 1(c) explaining the challenges with few examples. In contrast, background information is often not strong and easier to capture key object in other object detection problem. Hence, to stress on gun and suppress false positives, one solution could be strong local semantic representation of the scene based on the of gun.

CNN based classifiers are very efficient [6,7] in scene classification based on key objects. Till now most of the key object based classifiers are trained and tested on the images where key objects are occupied many pixels and mostly present in the center of the natural image. Unlike that, till now certain challenges are remain in gun (a small key object) based scene classification. The challenges regarding gun based scene classification are as follows:

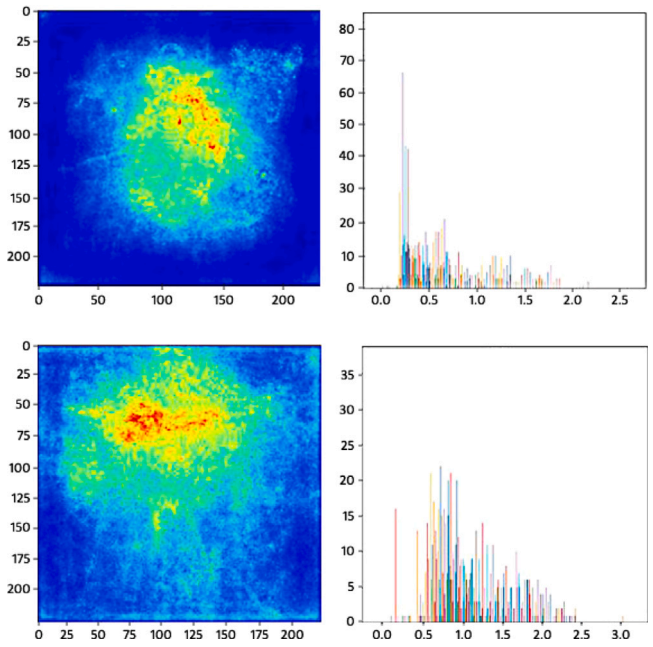

Fig. 2. (a) Heat map of natural image features collected from ImageNet Dataset. (b) corresponding graph showing key-object distribution. (c) Heat map of images with guns. (d) Corresponding graph showing variable distribution of gun in images.

- **Extraction of semantic based feature:** Gun based scene classification is difficult compared to the other object based scene classification. For example: in human or other object based scene classification the key object is condensed in the given scene and occupies many pixels. Hence, global semantic work in this case. Wherein, gun based classification gun occupies few pixels and the gun can be present in any position. Hence, global semantic will not be appropriate regarding classification of scene based on gun. Global semantics may lead incorrect representation of the gun in the scene. In this case, there is a scope of local semantic features to distinguish between a scene with and without gun. Current deep networks are weak in representing the local semantic of input images.
- **Loss of low and middle level features** During deep networks training, loss of low and middle level features cannot be avoided [8,9]. As mentioned earlier, for gun based scene classification local semantics is important. Hence, in this case we need to retain the low and middle level features through the training of deep networks [10,11].

Objectives: In the scope of this work, we designed a CNN architecture to tackle the challenges discussed previously. Hence, the objectives of this work are as follows:

- **Preserve the low and middle level features:** The proposed CNN model needs to preserve low and mid-level features, unlike traditional deep CNN models. In deep CNN models, as the depth increases, low and mid-level features tend to vanish. The proposed architecture also retains the gradient through each deeper layer, thereby addressing the gradient vanishing problem commonly encountered in deeper CNNs.
- **Critical representation of scene** Here, "critical" refers to the representation of scenes while considering the smallest objects present. The semantic features related to these small objects enable the classification of scenes based on the presence of a firearm. Therefore, the proposed architecture should effectively represent any scene.

The Proposed CNN architecture is based on the CNN architectural design of Qi Bi et al. [12]. The structure of the paper is also inspired by the work of Qi Bi et al. [12]. But there is no similarity regarding the objective and contribution. In the [12] they have been designed the CNN architecture for aerial scene classification, whereas we designed our architecture regarding the classification of scene based on the presence of Gun. For gun (small object) detection the proposed network used the backbone structure of Qi Bi et al. [12]. In [12], Dense block have been used as a backbone structure and afterwards a spatial attention module has been incorporated. Whereas, in the proposed network, the attention module has been used after each dense block and the attention module is not only comprise spatial attention it also do the channel attention. In [13], the importance of spatial and attention based module has been stated clearly though the application area is different. The proposed work also inspired from these two literature and design an attention module by incorporating both spatial and channel attention module. In the proposed architecture for better representation of the features an enhancement layer has also been included after the each attention module. In Section 4.1 (Ablation study), we also mentioned the effectiveness of these enhancement layer. At the end, before softmax classifier a global average pooling has been used for further enhancement of the features.

Contribution: This paper introduces a novel deep learning architecture that utilizes dense connections for the classification of scenes based on the presence of guns. The study emphasizes the challenges associated with classifying scenes containing small objects, such as guns, which can be difficult to detect due to their size and potential orientation in various scenes. The key contributions of this work are outlined as follows:

- **Channel-Spatial Attention Network:** We propose a channel-spatial attention network designed to enhance the detection of small objects within a scene. Given that guns are typically small objects that can appear in any orientation, our attention module is specifically crafted to effectively identify and highlight guns regardless of their position or orientation in the scene.
- **Innovative Use of Attention Modules:** Unlike existing attention-based networks, where the attention module is typically applied at the end of the backbone structure, our proposed architecture integrates the attention module after each dense block. This approach significantly improves the model's ability to consistently mark the presence of a gun throughout the input scene.
- **Enhanced Semantic Representation:** We introduce an enhancement layer following each attention module, which further refines the semantic representation of the scene with a focus on guns. This layer ensures that the model can accurately capture and represent the small but critical details that indicate the presence of a gun.
- **Comprehensive Experiments and Analysis:** We conducted extensive experiments on benchmark datasets to validate the effectiveness of our proposed architecture, GSNet. Additionally, we provide a thorough experimental analysis and an in-depth discussion to comprehensively evaluate the performance of GSNet, demonstrating its superiority in the task of gun-based scene classification.

2. Related work

Several research works reported for classification of input image into either gun category or non-gun category based on the presence of gun. In this section, we briefly described the classifier based methods for scene interpretation based on the presence of the small objects and also on the presence of the gun.

2.1. Classification of scene based on the presence of specific small object

Yu Chen et al. [14] describe the challenges for the classification of small sample target objects in the sky. To overcome the problem their proposed work used sparse auto encoder model to extract local features. For the global features they rely on the CNN. The extracted local features and target images were fed into the CNN for the accurate classification of the small sample target object in the sky. Paul Tresson et al. [15] proposed a hierarchal classification methodology for scene classification based on the presence of small objects like butterflies. In this method authors first trained a detector to detect object from the input image and then cropped the object area from the given image. Then the cropped object are classified with the trained classifier. Though the methodology is complex but its results are convincing. Small object detection is closely related to the classification of small object as backbone of any object detection methodology is an accurate classification algorithm. Likewise, Jiahe Zhang et al. [16] proposed a bar code (Small object) detection mechanism based on region proposal. They proposed a Region proposal network with classification layer which can be represented as a classification mechanism with extracted features. The difference is, the proposed method doing classification based on the region. Afterwards, CNN based detection algorithm is employed to detect the small object. Later, Jun Jia et al. [17] improved the model with an extension of distortion removal module. Distortion removal methodology is used to remove the geometric distortion according to regression parameters acquired from the previous step. The accurate position and distorted barcodes shape can be determined and corrected by this method. Jun Jia et al. [18] also used a weights pruning and recoding to reduce storage and memory overheads. Adnan Sharif et al. [19] aims to decoding of barcode automatically. For decoding a barcode, employed method required to detect the bar code which is indeed a small object. They used Dilated residual networks for the classification and localization of specific object as it will be available to capture both local and global features. They used iterative pruning compression technique to compress the model to reduce time complexity. In other work, Adnan Sharif et al. [20] extend the work by employing GAN model to trained the model on GAN generated blurred barcode. In addition to classifiers, several studies focus on small object detection using feature fusion techniques. Feature Pyramid Networks (FPN) [21] utilize multi-scale feature fusion to enhance object detection at various resolutions, improving the recognition of objects, particularly smaller ones. AFPN [22] (Asymptotic Feature Pyramid Network) builds on FPN by introducing asymptotic feature aggregation, further optimizing multi-scale detection. EfficientDet [23] refines this approach with BiFPN (Bidirectional Feature Pyramid Network), ensuring both efficiency and high accuracy. For text detection, CM-Net (Concentric Mask-based Text Detection) [24] specializes in identifying arbitrarily shaped text, while the Zoom Text Detector [25] employs a zoom-in mechanism to enhance small text recognition. PANet (Path Aggregation Network) [26] advances instance segmentation by integrating adaptive feature pooling and bottom-up path aggregation, refining segmentation performance. The Reinforcement Shrink-Mask [27] method further enhances text detection by utilizing reinforcement learning for precise mask refinement.

2.2. Classification of scene based on the presence of gun

One of the important works in the gun detection domain was performed by J. Lai et al. [5] on 2017. J. Lai et al. used googlenet to classify an input image either as positive or negative image based on the presence of the Gun. The limitation of this work is, a very simple video dataset have been used. Afterwards, to know the effectiveness and performance of deep learning methodologies for classification of images based on the presence of a gun, on 2018 S. Akcay et al. [28] implement several deep based classifier for classification of images/video with gun. In this work, the deep architectures were trained from scratch to obtain

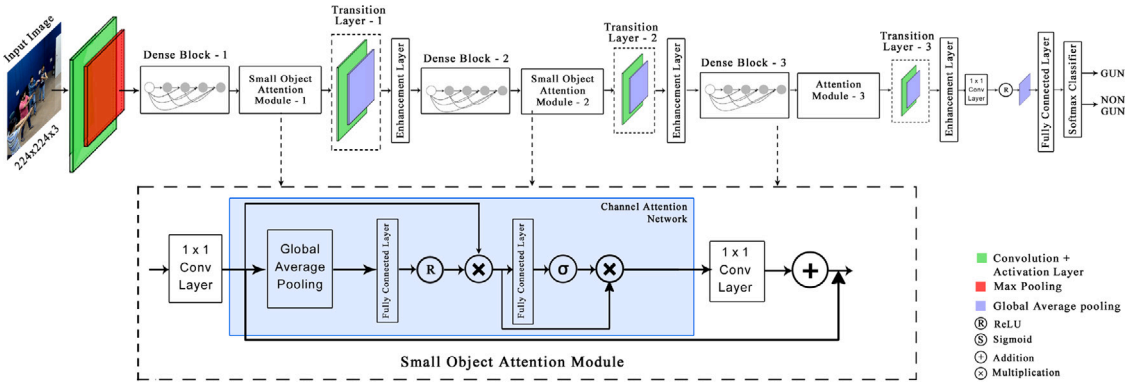

Fig. 3. Framework of the proposed CNN Architecture.

the correct weights specific to this task. As well as pre-trained weights are also tested and noted. A challenging dataset have been used in this work and the conclusion says that RESNET-50 performs better in comparison to the other networks.

Likewise, in [29]a comparison between the CNN classifier and conventional classifier has been drawn. And with obvious reason it has been concluded that deep based classifiers performed well compared to the conventional classifier. In [30], used specific region based features rather than the holistic feature to reduce the feature vector's size. To find out the Gun area specific features, YOLO has been used at the first hand for detection of gun. The proposed architecture overcomes the drawback of previous work for the sake of time and complexity. In this scope of work, the primary focus is the classification of the scene based on the presence of gun but without implementing any detection methodology. Currently, attention modules are implemented with the deep learning models to enhance specific features. As name implies attention modules are able to efficiently provide attention to the specific features either high level features or low level features. Recent days, attention modules have been applied in various research areas such as visual saliency prediction [31], visual attention models for human eye fixation [32], medical image prediction & segmentation [33], object detection [34], satellite image processing [12] etc.,.

2.3. Densenet in classification and detection of small object

DenseNet has emerged as a strong architecture for small object detection due to its ability to efficiently propagate features across layers, making it particularly useful in cases where fine-grained spatial details are crucial. In ship detection, for instance, Zhelin Li et al. [35] integrated DenseNet with YOLOv3 to enhance detection in maritime environments, where small ships often appear in cluttered backgrounds. DenseNet's dense connections allow better feature reuse, ensuring that low-level spatial features essential for small object recognition are preserved throughout the network. This approach improved accuracy while maintaining a lightweight structure suitable for real-time detection. Similarly, Xu Han et al. [36] applied DenseNet to Single Shot Detection (SSD) to enhance the detection of tiny targets, a common limitation in SSD due to its reliance on single-pass detection. By fusing DenseNet's low-level and high-level features, the model retained spatial resolution necessary for tiny object detection, such as small vehicles or distant objects in complex scenes. Additionally, Mingyang Pan et al. [37] employed DenseNet for multi-scale feature fusion in SSD, addressing scale variation issues that SSD traditionally faces. DenseNet's ability to propagate features across layers ensured that both local and global information were fused effectively, boosting detection performance, particularly for smaller objects. However, while DenseNet helps mitigate issues like vanishing gradients and poor feature retention, its dense connections can introduce computational complexity and increased memory usage, potentially impacting real-time applications. Nevertheless, DenseNet's strengths in retaining critical features and supporting multi-scale fusion make it a valuable tool for improving small object detection across diverse domains.

The limitations of these DenseNet-based small object detection methods include increased computational complexity due to dense connectivity, which can slow down real-time processing. Additionally, DenseNet's dense feature propagation leads to higher memory usage, posing challenges for deployment in resource-constrained environments. Finally, while these methods improve small object detection, they can struggle with detecting objects in highly occluded or extremely noisy backgrounds, limiting their robustness in certain real-world scenarios.

3. Proposed GSNet

3.1. Network overview

Fig. 3. is the pictorial representation of the proposed architecture GSNet, and Table 1 present the configuration of the network. The GSNet comprises four primary module, Dense block, transition Layer, Small Object Attention network (SAN) module and Classification Layer. Each dense block followed by a transition block and SAN block. The proposed architecture consist 3 dense block. After the last SAN module, the output feature map inputted to the classification layer.

3.2. Dense block

G. Huang et al. [38] proposed DenseNet architecture to achieve high accuracy in classification problem. In dense net, output of each convolutional layer are feed to the next convolution layers up to the last convolution layer of dense block. The dense net represents feature reuse throughout the network, remove the gradient vanishing problem and also reduce the loss of low and middle level features. Our GSNet is based on the modified Dense Network architecture. More specifically, the difference between the commonly use convolutional layer and a dense block can be perceived by the following equations. Where, Eq. (1) represents working of convolutional layer and Eq. (2) represents working of dense block.

$$x_l = H_l[x_{l-1}] \tag{1}$$

$$x_l = H_l[x_0, x_1, \ldots, x_{l-1}) \tag{2}$$

where, H is the composite function. A set of convolution layers (Conv), activation functions along with batch normalization (BN). The composite function can be comprised any number of convolution and other layers in any order.

Therefore, Eq. (1) represent that the output of $l_{th}-1$ layer composite function is inputted to the next l_{th} layer and most of the existing CNN architecture follows this arrangement. Whereas, Eq. (2) tells that

Table 1
Network structure of the proposed GSNet.

Layer	Output size	Proposed network
Initial Convolution	112 × 112	Kernel size: 7 × 7, stride: 2, num: 64
Pooling	56 × 56	Kernel size: 3 × 3, max pool, stride: 2, num: 64
Dense Block-1	56 × 56	$\begin{bmatrix} 1\times 1 conv \\ 3\times 3 conv \end{bmatrix} \times 3$
SAN Module - 1	56 × 56 1 × 1 1 × 1 1 × 1 56 × 56	Kernel size: 1 × 1, num: 64, Convolution Kernel size: 56 × 56, Global Average pooling Fully Connected Layer, num: 16 Fully Connected layer, num: 64 Kernel size: 1 × 1, num: 1
Transition Layer-1	56 × 56 28 × 28	Kernel size: 1 × 1 convolution Kernel size: 2 × 2, Average pool, stride: 2
Enhancement Layer	28 × 28 28 × 28	Kernel size: 1 × 1 convolution Kernel size: 1 × 1 convolution
Dense Block-2	28 × 28	$\begin{bmatrix} 1\times 1 conv \\ 3\times 3 conv \end{bmatrix} \times 3$
SAN Module - 2	56 × 56 1 × 1 1 × 1 1 × 1 56 × 56	Kernel size: 1 × 1, num: 64, Convolution Kernel size: 56 × 56, Global Average pooling Fully Connected Layer, num: 16 Fully Connected layer, num: 64 Kernel size: 1 × 1, num: 1
Transition Layer - 2	28 × 28 14 × 14	Kernel size: 1 × 1 convolution Kernel size: 2 × 2, Average pool, stride: 2
Enhancement Layer	14 × 14 14 × 14	Kernel size: 1 × 1 convolution Kernel size: 1 × 1 convolution
Dense Block-3	14 × 14	$\begin{bmatrix} 1\times 1 conv \\ 3\times 3 conv \end{bmatrix} \times 3$
SAN Module - 3	56 × 56 1 × 1 1 × 1 1 × 1 56 × 56	Kernel size: 1 × 1, num: 64, Convolution Kernel size: 56 × 56, Global Average pooling Fully Connected Layer, num: 16 Fully Connected layer, num: 64 Kernel size: 1 × 1, num: 1
Transition Layer - 3	14 × 14 7 × 7	Kernel size: 1 × 1 convolution Kernel size: 2 × 2, Average pool, stride: 2
Enhanced Classification Layer	7 × 7 1 × 1	Kernel size: 1 × 1 convolution Kernel size: 7 × 7 global average pool Fully Connected Layer, Softmax Layer

output of each layer's composite function $x_0, x_1, \ldots, x_{l-1}$ are inputted to the l_{th} layer. The prime advantage of this arrangement is the flow of information through each layer.

In denseNet [38], each dense block are organized as BN → Activation Function → 1 × 1 Conv → BN → Activation Function → 3 × 3 Conv. In Dense net, Rectified Linear Unit (ReLU) used as activation function. The discussed dense net contains 4 dense block and 3 transition layers. Each dense block followed by one transition layer. Transition layer consist of one 1 × 1 convolution layer followed by average pooling layer. Transition layer is introduced to reduce the size of output feature map to further increase the compactness of the network. A compression rate θ is used in transition layer to quantitatively measure the reduction. Intuitively, if h is the number of output feature maps of any transition layer then it can be represented by the following equation-

$$h = \theta \times k \tag{3}$$

Where, k is the number of input feature maps and $0 \geq \theta \leq 1$. Note that h is the output feature map of the 1 × 1 convolution layer of transition layer. In DenseNet [38], θ is set to 0.5 to reduce redundancy of feature.

Why Dense net

With significant advancements in convolutional neural networks (CNNs) such as VGGNet [1], GoogleNet [2], Inception-V4 [3], ResNet [4], and DenseNet [5], these architectures have become increasingly popular as classifiers. DenseNet, in particular, offers unique features that enhance its ability to detect small objects, which is why we chose to use DenseNet in this work. DenseNet aims to improve upon the performance of ResNet [cite DenseNet paper], and both architectures share the commonality of feature reuse. However, the primary difference lies in how they aggregate features: while ResNet combines them through summation, DenseNet does so through concatenation.

Research [6] suggests that ResNet's use of summation can increase the risk of "washing out" important feature maps. On the other hand, as demonstrated in multiple studies [7,8,9], networks with multiple receptive fields can capture visual cues effectively across various scales. DenseNet addresses this challenge by preserving prior feature maps through feature concatenation, as shown in [relevant equation from DenseNet paper]. Since our task involves classifying scenes that contain firearms typically small objects preserving detailed information at each layer is crucial.

Moreover, unlike other networks, DenseNet provides direct access to the gradients from the loss function and the original input, which helps mitigate the issue of vanishing gradients due to improved information flow. Therefore, we selected DenseNet as the backbone of our architecture. To further enhance DenseNet's classification performance, we introduced additional components such as an attention module, enhancement layer, and modified classification layer.

In the proposed architecture, dense block and transition layer has been used but with different arrangement. The difference between Dense net and GSNet are as follows-

- In GSNet network, 3 dense block and 3 transition layer has been used instead of 4 dense blocks. In dense net, output of last dense block has been inputted directly to the softmax layer. Unlike to that in GSNet each dense block followed by a transition layer.

- In GSNet, θ is set to 1 for highlighting semantics in each feature map. This different arrangement retain minute enhanced local features for highlighting. Afterwards, on the highlighted feature map average pooling has been implemented.

Considering the objectives of this work, there are several distinct characteristics of the proposed network setup-

- **Reuse of Features-** Dense blocks make reuse of features of all the layers. Features from the shallower layers losses with the increase of the depth of layers. In addition, feature map obtained from one layer is inputted to the next layer, hence the next layer can have the access of the abstract information outputted from the previous layer. To overcome this lacking, residual networks are proposed. Considering the Eqs. (1) and (2), Residual connection network can be expressed by the following equation-

$$x_l = H_{(l-1)} + x_{(l-1)} \tag{4}$$

  From the equation, it is evident that residual networks preserve the information of previous layers through a additive identity transformation. After several experimentation, it has been noted that few layer contribute very little & hence randomly few layers were dropped. Which in turn make the residual CNN networks works like recurrent CNN networks. Another problem with residual network is use of extensive number of parameters. Dense block has several advantages over residual CNN networks. Such as, in dense block information preservation is more efficient, there is clear differentiation between the information that is added to the network and the preserved information. Furthermore, dense blocks use fewer number of filters and convolution layers which reduces the number of parameters.
- **Fewer convolutional layers than DenseNet-** Proposed network using dense blocks, but with fewer convolution layers than the DenseNet [38]. In original DenseNet, there are more than 100 convolution layer with 10 composition function. In contrary, GSNet comprise of 20 convolution layer with three composite function in each dense block. In DenseNet, 4 dense block and three convolution layers, whereas in GSNet last there are 3 dense blocks instead of 4. Hence, GSNet represent features in more orderly manner and parameters are reduced significantly.
- **Use of full channels in Transition Layer-** Considering the Eq. (3), θ is a compression rate which is used to control the number of channels in the output feature map. In original DenseNet, θ is set to 0.5, to reduce the channels and make the network more compact. In GSNet, θ is set to 1 to use maximum number of channels of feature maps. As the parameters and number of feature maps are reduced significantly in GSNet, θ is set to 1. The arrangement enhance the local semantics of the input image.

3.3. Small object attention module (SAN module)

GSNet utilizes a Small Object Attention Module showed in Fig. 5 after each dense block and its subsequent transition block (Discussed in Section 3.2) to assign greater weight to small objects, enhancing scene understanding. As the proposed work is meant to classify input scene based on the presence of gun, weighted attention is appropriate procedure to implement along with the backbone network. Hence, in this scope of work an attention module has been proposed. Proposed Attention module (SAN Module) consists of two sub modules, spatial attention module and channel-wise attention module. In the following subsections, each sub-module is discussed individually, followed by an explanation of how the two sub-modules are interconnected to form the Small Object Attention Module.

3.3.1. Spatial attention

To impose weight based on the spatial arrangement of the output feature map $P_{i,j}$, in this work a composite function has been used. The composite function comprises with-

1×1 convolution → Activation function → 1×1 convolution → Activation function

The attention weight $T_{i,j}$ is calculated via:

$$T_{i,j} = (Softmax(W_2 \ \ tanh(W_1 P_{i,j} + b_1) + b_2)) \tag{5}$$

Here, W_2 and W_1 represent the trainable weight matrices, and b_1 and b_2 are the corresponding bias terms. More specifically, the first 1×1 convolution, represented as $tanh(W_1 P_{i,j})$ applies the tanh activation function, while the second 1×1 convolution, parameterized by W_2, uses the softmax activation function.

3.3.2. Channel attention

Spatial attention module are not capable to extract semantic information by considering channels. During calculation of spatial attention weights each channels are considered and do not consider information from key channels. Suppose, F is the set of output feature maps generated by any random convolution layer. Specifically, F is of size of $h \times w \times c$, we also can define $F = F_1, F_2, \ldots, F_n$ where, n is the number of channels and size of each F_i is $h \times w$ for $i = 1, 2, \ldots n$. Among these F_i's there may be few channels which represent the scene more meaningfully, those channels are called key channels. Hence extraction of more channel wise information may make the network to learn distinctive features. To deal with these recently, channel attention modules has been introduced in various research [39–42]. Inspired by this, in GSNet we develop a channel attention network for mining the key channels and fuse channel attention weights with the spatial attention weights. Channel attention networks uses global information to selectively emphasize informative features and suppress lesser ones.

In this scope of work, channel attention weights are obtained by a composite function $S(.)$. The composite function contains

Global Average Pooling → Conv1 → Activation1 → Mult → Conv2 → Activation2 → Mult.

Here, Conv1 and Conv2 represent 1×1 convolutions, while 'Mult' denotes pixel-wise multiplication of feature maps. As shown in Fig. 3, which illustrates the attention network, the output of the Global Average Pooling (GAP) layer is multiplied with the output of each convolution layer. This process ultimately generates the final channel attention weights.

3.3.3. Spatial and channel attention arrangement in SAN module

In previous sections we have discussed about the two components of SAN module, channel attention and spatial attention. We are accommodating both the module in one block to create proposed Small Object Attention Module. Now the question is- how both the block were arranged in one block? In the proposed SAN (Spatial and Channel Attention) module, we integrate the channel attention block within the spatial attention block to enhance feature representation. After the initial 1×1 convolution in the spatial attention block, the channel attention module is applied. The attention weights generated by the channel attention block are then multiplied with the input feature map, refining the feature maps before continuing with the spatial attention layers. This arrangement allows the calculation of spatial attention weights on the channel-refined features, improving the network's focus on both spatial and channel-wise information. The process can be described step by step mathematically, considering the input feature map I of size $H \times W \times C$ as follows:

On the feature map I, the first step of the spatial attention block consists of applying a 1×1 convolution followed by an activation function. Mathematically, this can be expressed as:

$$I_{conv1} = RELU(w_1 * I + b) \tag{6}$$

Here, I_{conv1} is the output feature map after applying a 1×1 convolution with a ReLU activation function. W_1 represents the weight matrix of size $1 \times 1 \times c \times c_1$, while b_1 is the bias term. The symbol * denotes the convolution operation, and ReLU is the element-wise activation function applied to the result.

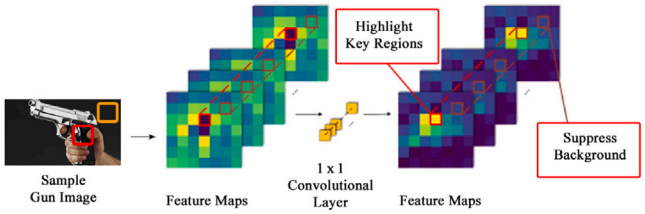

Fig. 4. Demonstration about the importance of 1 × 1 convolution layer for highlighting features.

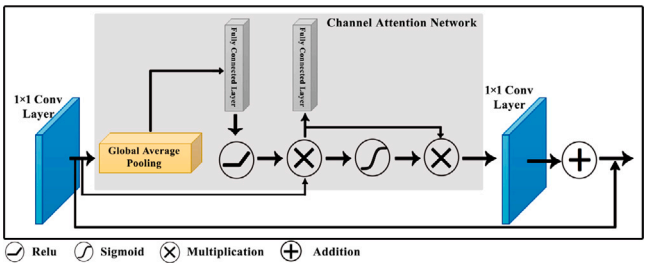

Fig. 5. Demonstration of proposed Small Attention Module (SAN Module).

Next, Global Average Pooling (GAP) is applied to the output I_{conv1} within the channel attention block. This operation reduces the spatial dimensions, summarizing each channel into a single value. The resulting output, I_{gap}, has a size of $1 \times 1 \times c_1$. This process can be described by the following equation:

$$I_{gap} = GAP(I_{conv1}) \tag{7}$$

Here, GAP represents the global average pooling operation, which averages the spatial dimensions of the input feature map for each channel. After the Global Average Pooling (GAP) operation, the channel attention module consists of two fully connected layers, each involving two sequential multiplication operations. The process can be mathematically expressed through a series of equations as follows, with each operation explained in detail following the equations. For the first fully connected layer:

$$W^{(1)}_{channel} = \sigma(W^{(1)}_{2}.Relu(W^{(1)}_{3}.I_{gap} + b^{(1)}_{3}) + b^{(1)}_{2})$$

Where, $W^{(1)}$ and $W^{(2)}$ are trainable weight matrices, $b^{(2)}_{2}$ and $b^{(3)}_{2}$ are bias terms and σ is indicating sigmoid activation function. Now the weight matrix will be multiplied with the output of the first convolution (refer to equation):

$$I^{(1)}_{weighted} = W^{(1)}_{channel} \otimes I_{conv1} \tag{8}$$

Here, I_{conv1} represents the refined feature map, enhanced by the channel attention mechanism. Similarly, subsequent convolution and multiplication operations apply the channel attention mechanism after each layer. In this work, we employ two fully connected layers, and the refined feature map can be obtained using the following equation:

$$W^{(2)}_{channel} = \sigma(W^{(1)}_{4}.Relu(W^{(1)}_{5}.I^{(1)} + b^{(1)}_{4}) + b^{(1)}_{5}) \tag{9}$$

$$I^{(2)}_{weighted} = W^{(2)}_{channel} \otimes I^{(1)_{weighted}} \tag{10}$$

Following the channel attention block, a final 1 × 1 convolution is applied to compute the spatial attention map, as described by the following equation:

$$W_{spatial} = \sigma(W_6 * I^{(2)}_{weighted} + b_6) \tag{11}$$

The final refined feature map is obtained by multiplying the spatial attention weights with the channel-refined feature map. This process is mathematically expressed as follows:

$$I_{final} = W_{spatial} \otimes I^{(1)}_{weighted}. \tag{12}$$

3.3.4. Residual attention block

From the earlier discussion, the attention block refines the feature maps using both spatial and channel attention mechanisms. However, a known issue with this bottom-up approach is the potential loss of important information after each attention block. Specifically, the original input feature map, denoted as $I \in \mathbb{R}^{H \times W \times C}$ can be lost after the application of the attention mechanism. This is problematic, especially in deeper architectures where preserving the original feature representation is crucial for effective feature extraction. To address this issue, we propose incorporating a Residual Attention Block within the SAN (Spatial Attention Network) module. This block introduces a residual connection that preserves the original input feature map while allowing the attention mechanism to refine the feature representation. The residual connection operates through the identity function, ensuring that the attention block does not disrupt the continuity of the original feature information as the network progresses. The residual attention mechanism can be mathematically expressed as follows:

$$I_{final} = W_{spatial} \otimes I^{(1)}_{weighted} + I_{conv1}. \tag{13}$$

This generalized formulation ensures that the original features are maintained even as the attention weights enhance the discriminative power of the network. The residual attention mechanism thus allows the network to selectively focus on critical areas of the feature map without losing the essential structural and contextual information encoded in the original map.

The integration of the residual attention block into the SAN module allows the network to continue its bottom-up convolutional process without being disrupted by the attention mechanism. The residual connection ensures that the attention mechanism acts as a refinement rather than a transformation that alters or erases important features.

The output feature map I_{final} from the residual attention block is subsequently passed to the transition layer, which prepares the feature map for the next dense block. This transition layer ensures that the training process continues smoothly in the subsequent dense block by maintaining a balance between refined attention-based features and original feature information.

Thus, by incorporating the residual attention mechanism, the SAN module achieves better feature preservation, especially for complex scenes where small objects or fine details play a critical role. This approach also mitigates the risk of vanishing gradients or information loss in deep neural networks, ultimately leading to improved performance.

3.4. Enhancement layer

In GSNet, two 1 × 1 convolution is used to enhance the local semantics in the output feature map of each transition layer. After each transition layer, output feature map inputted in enhancement layer. The enhancement block comprises of two 1 × 1 convolution. Fig. 4 showed, enhancement of local semantics using 1 × 1 convolution. Further improvement in GSNet used two subsequent 1 × 1 convolution layer. 1 × 1 convolution layer imposed weight to the local patches which are relevant to the scene. The input size and output size of output feature will be same, but it will highlight key region in the feature map.

3.5. Enhanced classification layer

In general, feature map obtained after set of composite functions, it is send to the softmax classifier. Softmax Classifier usually calculate the probability of the ground truth level. The traditional softmax classifier converts the network's final feature map into class probabilities by applying a linear transformation (fully connected layer) followed by the softmax activation.

As we are considering the presence of gun for representation of the input scene and the primary challenge of the problem is high inter class variability and low intra class variability. Hence, it required

strong representation of local semantics at the classification layer. In this proposed network, we used few components to enhance the classification layer. We define a composite function for classification layer comprising a 1 × 1 convolution, afterwards a ReLU activation function and 7 × 7 global average pooling. Output of the global average pooling inputted to a fully convolution layer and then to a softmax layer. More specifically, consider, $W_1^{1\times1\times h}$, $b_1^{1\times1\times h}$, $W_2^{7\times7\times h}$ are weight matrix & bias matrix of the additional 1 × 1 convolution layer, weight matrix of 7 × 7 global average pooling respectively and ReLU represent relu activation function. Thus the output of enhanced convolutional layer can be expressed as follows-

$$H_f = W_2^{7\times7\times h}(ReLU(W_1^{1\times1\times h}H^{7\times7\times h} + b_1^{1\times1\times h})) \quad (14)$$

H_f is the output of the enhanced classification layer and H^{7x7xh} is the input of enhanced classification layer. The output H_f is inputted to the fully convolution layer which in turn pass into the final softmax layer. Now by considering the W_f^{1x1xl} and b_f^{1x1xl} is the weight matrix and bias matrix of fully convolution layer. Hence the input of the softmax layer will be:

$$H_f = W_f^{1\times1\times l}(W_2^{7\times7\times h}(ReLU(W_1^{1\times1\times h}H^{7\times7\times h} + b_1^{1\times1\times h}))) + b_f^{1\times1\times l} \quad (15)$$

There are certain advantages of this enhancement. Feature Refinement Before Pooling: The traditional softmax directly pools and flattens the feature map for classification, but your network first refines the features through a 1 × 1 convolution and non-linearity (ReLU), ensuring that small yet critical features are highlighted before pooling.

Localized Feature Retention: The 1 × 1 convolution retains spatial information across the image, crucial for small object detection tasks, whereas the traditional softmax classifier often loses this localized information when flattening the entire feature map into a single vector.

Global Context Awareness: The global average pooling layer ensures that the network maintains global context, particularly important in tasks where both the object and the surrounding scene are important for classification (e.g., a gun in a crowded scene).

Improved Handling of Small Objects: By refining features through additional layers, your proposed classification layer is better suited to tasks like small object detection in complex scenes, where subtle and localized features must be retained before classification.

3.6. Relation to existing attention modules

CBAM [42] was the first network introduced with feature enhancement capabilities, using channel and spatial attention to refine feature maps. In CBAM, attention is applied sequentially to improve feature representation. Similarly, BAM [43] applies attention at the bottleneck layers to enhance feature learning, while SENet [41] focuses exclusively on channel recalibration. Additionally, HRNet [44] employs parallel convolutional streams to maintain high resolution feature representations throughout the network.

In contrast, the proposed methodology integrates SAN Modules within Dense Blocks, enabling deeper feature enhancement across multiple stages. Enhancement layers are incorporated to further refine the features before they are passed to the next stage, ensuring a more effective feature representation. Unlike BAM, which applies attention only at bottleneck layers, SAN Modules are distributed throughout the network, including Dense Blocks and Enhancement Layers, allowing for multi-scale feature extraction. Furthermore, by combining channel attention with spatial feature refinement, the proposed methodology extends beyond SENet's single-stage channel recalibration, making the feature extraction process more comprehensive. Instead of relying on parallel convolutional streams like HRNet, feature resolution is refined through Enhancement Layers, optimizing high-resolution learning more efficiently.

4. Experimental results

4.1. Database description

In this scope of work, TUVD-CSA Dataset has been used to implement the proposed architecture as well as competing methodologies. The Dataset is designed by us and the detail description can be found in [48] and it is available for research community in (http://www.mkbhowmik.in). Quality of the images/videos is key factor for efficient implementation of any computer vision systems [49–54]. If the images are not of good quality then it requires to implement pre processing methodologies to enhance the quality. The dataset used in this work has been captured with high end camera hence we are not applying any enhancement algorithm for up-gradation of the images. Next, we segregate the dataset in two classes:

- Class 1- it is composed of the images either having person who is carrying gun or only having the gun.
- Class 2- it is composed of the images without having any trail of gun. These images may contain objects alike guns such as bottle, mobile phone etc.,

The Database composed of 65 indoor videos, 60 outdoor videos and 25 videos are downloaded from the web. In total, there are 150 nos. of videos with more than 5 lakh frames. The idea of including the downloaded videos is to increase the complexity of the dataset. The remaining 125 videos are captured in Laboratory, Classroom for indoor condition and in parking places, building premises, corridors, garden, open fields, different crossings (3 way, 4 way) for outdoor conditions. Note that all the videos are captured in the Tripura University (A Central University), Suryamaninagar, India. A Nikkon D5 100 camera with 30 fps (Frame per second) is used for the capturing of the videos. The camera is positioned on a tripod at an angle of 60°. Note that the height of the tripod is 9ft – 12ft from the ground. Students of Tripura university volunteered to annotate actions of a suspected person with guns in hand. The dataset demonstrate several features of dataset such as Effects of illumination change, occlusion, rotation, pan, tilt, scaling of gun and has been shown in Fig. 6.

Aim of the proposed CNN architecture is to classify input scene into either positive scene (with gun) or negative scene (without gun). Hence, two categories of dataset are required to learn the features for scenes with gun. By following the same procedure, we have divided the dataset into two prime categories, Class-A and Class-B. In class A, images with firearm (carrying by persons) are present and in Class B, images without firearm are present. The images are representing several challenges likewise lighting change (from indoor to outdoor), illumination change, partial occlusion etc.,

Data augmentation is an important step to be followed before applying CNN architecture. Data augmentation increase the amount of data and also increase the variation in data. In this scope of work, we transform data to simulate realistic views of the object to be detected. Before augmentation, all the images are resized to 640 × 480 pixels. The following augmentation methodologies have been applied:

- Increasing and decreasing the brightness by (20%) in order to simulate the luminosity change. This approach replicates diverse lighting environments, including indoor, outdoor, and low-light conditions, ensuring the model accurately detects weapons across varying brightness levels without being dependent on a specific illumination setting.
- Flipping and rotations (±10°, ±20°, ±30°) to create the different canonical views of the object. This process generates various weapon orientations, mimicking different angles a gun may appear in real-world scenarios, and strengthens the model's robustness to perspective changes, improving detection reliability.
- Cropping the gun region from the whole frame to simulate different shapes and colors of the gun. That helps the model focus on local weapon features rather than background context.

Table 2
Statistics of publicly available dataset.

Database	Publication year	No. of images/videos	Image format	Database type	Resolution	Weapon type
IMFDB [45]	2014	450,000[a]	.jpg	Image	Variable size	Gun
Knives Images Database [46]	2015	12,899	.bmp	Image	100 × 100	Knife
Gun Movies Database [46]	2013	7 Videos	.mp4	Video	640 × 480	Pistol
Dataset of R. Olmos et. al [4]	2018	9261	.jpg	Image	640 × 480	Gun
Dataset of D. Ramerio et. al [30]	2019	17,684	.jpg	Image	224 × 224	Gun
ITU firearm dataset [47]	2019	10,973	.jpg	Image	480 × 800	Gun

[a] Approx total Images, NP = Not Provided.

Fig. 6. Different features of the dataset like gun placed near to the camera, gun placed far from the camera, indoor control scenarios, outdoor scenarios, complex background, illumination changes, mass firing etc.,.

We also implement the proposed approach on existing datasets. Table 2 lists the publicly available datasets along with the proposed dataset, **TUVD-CSA** indicating the publication year, no. of negative and positive images in each dataset, weapon type, image resolution, type of files (video or image file). Table 2 also indicate the number of training and testing data ratio.

4.2. Experimental setup

Model initialization: To start training the proposed network is required to setup the weight matrices for different convolution and activation layers. After training procedure, we will obtain the final weight matrices. During the start of training procedure, the weight matrices are initialized randomly. There is another matrix, bias matrix required for train the CNN network. The bias matrix is set to be 0.001 to start training.

Training Procedure: Training procedure conducted in two stages. During first stage, the network trained with the learning rate of 0.001. Afterwards in next subsequent stages learning rate is decreases by 10 and the reduction will continue until the completion of the training. For different dataset, the two stages of training is different. For TUVD-CSA dataset after 150 epoch the reduction of learning rate is implemented. The same approach is followed for IMFDB and Dataset of D. Ramerio et al. dataset. For the other datasets having $< 13,000$ images/frames

Table 3
Classification accuracy of the proposed and other methods on TUVD-CSA dataset.

Methods	Training ratio	
	50%	80%
Fisher's LDA	59%	69%
SVM	67.7%	70%
VGGNet	85.4%	88.90%
ZFNet	84%	87%
GoogleNet	88.45%	90%
ResNet	87%	89.76%
DenseNet	90%	92%
PA	94%	97%
PA + Attention module	95%	97%

after 50 epochs the learning rate started to decreasing by 10. We used L^2 normalization with parameter 5×10^{-6} to avoid the over-fitting problem and we allowed 20% drop out during the training process.

Other implementation Details: As the proposed network is a 2 class classifier, the filter number in the final convolutional layer is 2 as per the number of scene category. The filter numbers of the final convolutional layer are different from the rest of the convolutional layer based on the number of scene category. Hence we specifically mentioned it.

All our algorithms are developed via Python under the TensorFlow framework. All the experiments were implemented on a work station with Xeon(R) E5-2630 v3 CPU and 64 GB memory.A TitanXP GPU was also used for acceleration.

4.3. Results and comparison

Several evaluation protocols are used to present the results of proposed network on the mentioned datasets. The results are compared with the other state-of-the-art CNN networks used for scene classification based on gun. Results of some baseline networks are also compared. Baseline networks and state-of-the-art CNN networks mentioned in this work all are implemented on the proposed dataset and the other dataset. We then demonstrate the classification accuracy on each challenging category to further investigate the performance. Finally, the model sizes and the prediction time per image is compared with other baseline CNNs.

4.3.1. TUVD-CSA dataset

Table 3 presents the results of the proposed Network and other methods (baseline & state-of-the-art methods) on the proposed dataset, TUVD-CSA. We trained the networks along with the proposed network with both 50% and 80% training ratio. Here, the training ratio referred to the ratio of training data and testing data. From the Table, following can be observed-

- Under the both training ratio, proposed network outperforms the baseline methods and state-of-the-art methods. The DenseNet architecture, backbone of our GSNet also performed well compared to the other networks.

- Network with the attention network (GSNet without SAN module) outperforms baseline networks and state-of-the-art methods but attain significantly low accuracy compared to the proposed network with SAN module.
- A difference between the results of 50% and 80% training ratio can be observed. Under the training ratio of 80% results are significantly improved compared to the results listed under the training ratio of 50% and it is expected as in first case more number of training images.

The performance of the GSNet on TUVD-CSA dataset can be explained by the following aspects:

- We can observed from above discussions that there are very few works attempting to classify scene based on presence of classification using deep classification network. Few research works has done extensive analysis on the performance of the baseline CNN networks on classification of scene with gun and scene without gun. These baseline methods tend to preserve global semantics whereas proposed CNN network capable of preserving features from different level and able to highlight local semantics of the input scene. The performance of proposed network without SAN module (GSNet without SAN module) represent this fact effectively.
- GSNet with Channel spatial attention module in proposed arrangement able to extract more discriminating features by imposing weights to significant local semantics. Hence, GSNet with channel spatial attention network outperformed other networks.

Moreover, Fig. 7 showed the classification accuracy of proposed network on several challenging conditions described in [55]. These results are shown under the both training ratio 50% and 80%.

Classification accuracy of baseline architecture effected by few challenging condition, such as sudden illumination change in the scene, partial occlusions of gun in scene. Comparably, proposed network performed better in these challenging situations. For example, GSNet achieve above 95% accuracy in the scenes having sudden illumination change, partial occlusion, mass firing etc., during 50% training ratio. Specifically, it can be concluded that our proposed network able preserve local spatial semantics as in partial occlusion, mass firing scenes global semantics are of no use. In most of the scenes, local semantics are important as guns or weapons are very small compared to the other objects. Hence, local semantics is of more importance than global semantics.

From the results it can be noted that in full occlusion of guns no methods are able to attain fair accuracy.

4.3.2. IMFDB

IMFDB is a dataset that are collected from the internet. We organize the dataset as per the challenging conditions which are mentioned in Table 4. It presents the results of baseline methods, state-of-the-art methods along with the GSNet. It can be seen that:

- Under the training ratio of 50%, proposed network (GSNet) with attention module outperforms compared to the state-of-the-art methods and baseline networks. In addition, it achieves highest accuracy in different challenging scene.
- Different aspect of proposed network such as, proposed network without attention module, proposed network without enhancement layer, and proposed network without enhanced classification layer also perform competitively with the state-of the-art methods.
- Compared to the baseline networks, state-of-the-art methods performed significantly well under the both training ratio of 50% & 80%.

the result findings can be explained as follows:

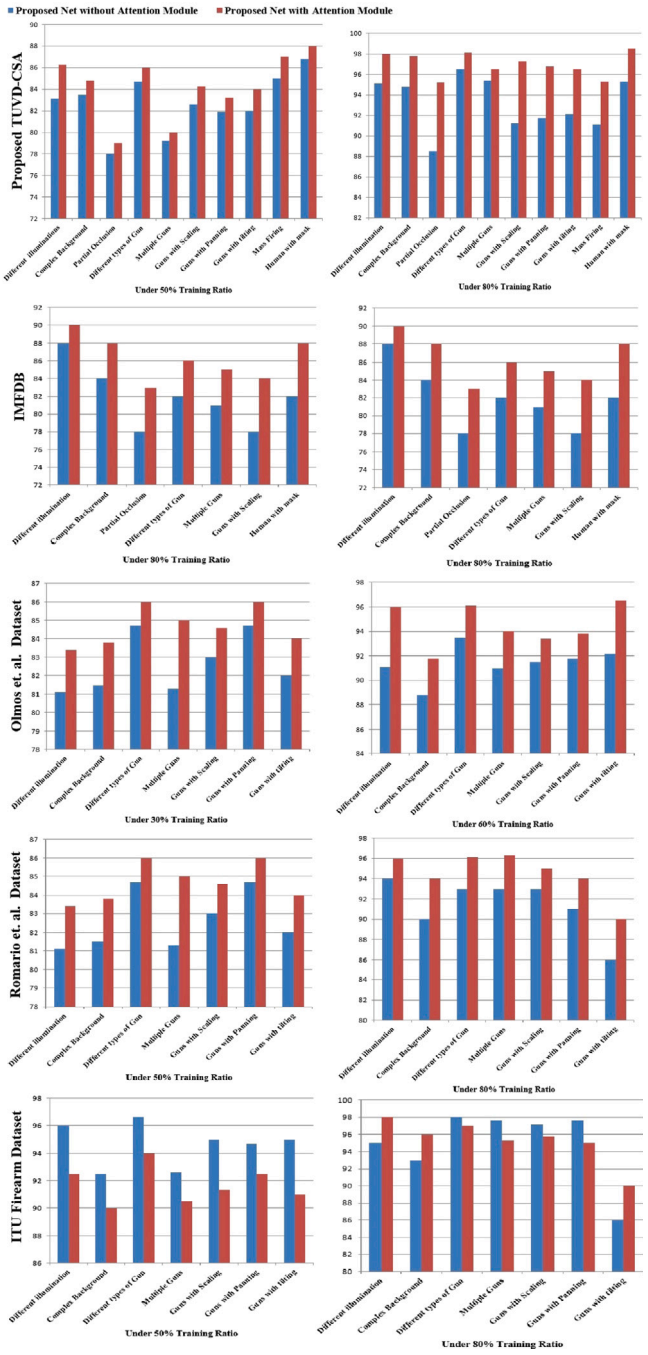

Fig. 7. Performance of the GSNet on the proposed dataset at 50% training ratio.

- IMFDB Dataset is also contains challenging conditions, where there is difficulty in distinguishing scenes in-terms of presence of guns. Therefore, IMFDB is also qualified as TUVD-CSA to validate a GSNet. It also shows that state-of-the-art method's improvements are significant compared to the baseline methods.
- Similar to the TUVD-CSA Dataset, in IMFDB dataset too performance of GSNet showed the significant effect of attention module and enhancement module. Proposed Network with attention modules outperforms proposed network without attention module (DCNN). In addition, the proposed model with both attention module and enhancement module outperforms all variants of proposed network.

Table 4
Classification accuracy of the proposed and other methods on IMFDB dataset.

Methods	Training ratio	
	50%	80%
Fisher's LDA	51%	53.2%
SVM	57.4%	67%
VGGNet	82%	86.90%
ZFNet	82.3%	85.9%
GoogleNet	89.5%	91%
ResNet	89.8%	91.2%
DenseNet	92%	94%
PA	95.8%	96.6%
PA + Attention module	96%	98.4%

Table 5
Classification accuracy of the proposed and other methods on gun movies dataset.

Methods	Training ratio	
	50%	80%
Fisher's LDA	70%	72%
SVM	87.3%	88%
VGGNet	91.1%	91.5%
ZFNet	92.4%	92.8%
GoogleNet	92.7%	93%
ResNet	92.6%	92.9%
DenseNet	95.8%	96%
PA	97.2%	98.8%
PA + Attention module	97.4%	98.4%

Fig. 7 showed classification accuracy of proposed network in each challenging scenes. Proposed Network performed well in each challenging situation even in the hard distinguishing scenes. Hard distinguishing scenes are referred to the scenes where gun is present not in usual condition or in a position which is not distinguishable. Proposed network is able to represent the key local semantics for distinguishing scenes with and without gun.

However, it is noticeable that in both the dataset, under training ratio of 50% proposed network performed well than under the training ratio 80%. It can be explained by the fact that proposed network preserve the shallower features such as texture, shape. The shallower features able to highlights key local regions and that is why with more training samples performance of proposed network decrease drastically.

4.3.3. Gun movies and knives image dataset

As discussed above, Gun movies dataset and knives image dataset is the most simple dataset and also contains less number of samples. The results of compared methods are tabulated in Table 5. The Results can be seen as follows:

- Under both the 80% and 50% training ratio, most of the methods performed well. In this dataset, no significant increase or decrease of proposed network has been seen. Knives are
- Even baseline methods also performed competitively to the improved State-of-the-art methods under both the training ratio.
- There is no noticeable difference of results can be observed under the 50% and 80% training ratio.

These results can be expressed as follows:

- As the dataset is simple with one gun and with simple & clear background networks do not required much time to saturate. Also not required more training sample to saturate.
- In this dataset, baseline methods are competitive to the proposed network and state-of-the-art methods which in turn over-fit the proposed and state-of-the-art methods.
- Results are not considerable as the dataset does not combat with real time scenarios. It can be noted that improvements of baseline networks is required to successfully classify input scenario based on the presence of gun.

Likewise the previous dataset, we organize the dataset as per the challenging scenarios. Fig. 6 presents the classification accuracy of proposed network for each challenge category under the training ratio of 50% and 80%. It can be noticed that the proposed Network achieves a fair classification accuracy results on all baseline challenging conditions (Refer to the Section 1 for challenging situation).

Moreover, proposed network with spatial attention module or with channel attention module brings notable improvement on each challenging conditions for gun based classification such as 'where the guns are capturing small number of pixels' such as crowded areas, with more number of false positives. The reason can be expressed as the key local semantics are taken into account.

4.3.4. Dataset of R. Olmos et al.

Table 6 showing the results of GSNet and baseline methods on the dataset designed by R. Olmos et al. The Dataset is designed by collecting images from the internet. The dataset comprises of 4 sub dataset namely: Dataset-1, Dataset-2, Dataset-3 and Dataset-4. The datasets are different in the aspect of number of classes and gun types. Results are evaluated under 30% and 60% training ratio as methods are saturated faster for lesser number of samples. The Results can be seen as follows:

- GSNet outperforms in all the sub-datasets compared to the state-of-the-art and baseline methods under both the training ratio. In Table 9 shows the average results of methods on the sub-datasets. Comparing to the baseline methods state-of-the-art methods are more competitive as per the results.
- Baseline methods failed miserably in Dataset-1 and Dataset-2 which reflected on average result of baseline methods. It is noticed that baseline methods perform fairly in Dataset-3 and Dataset-4. State-of-the-art methods and proposed network performed better than baseline methods which in-turn signify the improvement of the baseline methods in classification of challenging dataset.
- State-of-the-art methods and GSNet not perform expectantly in Dataset-1 and Dataset-2 under the training ratio of 20%

The results can be explained as follows:

- Reason behind the worst performance of baseline methods on Dataset-1 and Dataset-2 is the complexity of Dataset-1 and Dataset-2. Dataset-1 and Dataset-2 consist of several type of guns and huge number of false positives. As the baseline methods are represent the input scenes in terms of global features and hence cannot distinguish between the guns and non guns.
- GSNet wherein captures local semantics and due to the attention module it highlight the key local semantics. That is why GSNet performed well in Database-1 and Database-2 and hence average accuracy is also fair.
- Under 60% training ratio accuracy of all methods goes downwards and this can be expressed that CNN networks saturate near to the accurate weight matrices if it is trained on huge dataset. Each dataset contain comparable lesser number of samples and hence under training ratio of 20% accuracy of methods lowering down.

Fig. 7 showing the average classification accuracy of the GSNet in different challenging scenarios of the dataset under both the training ratio of 20% and 60%. It can be noted that, in this dataset scenarios of occluded guns are not available. In other scenarios GSNet able to attained fair accuracy. Under the training ratio of 20% performance of GSNet decreases significantly. The reason has been discussed previously.

The dataset contains a large number of images containing false positives. GSNet able to distinguish between images with false positives and image with gun. Moreover, we also seen that GSNet with attention module attain highest accuracy in this case.

Table 6
Classification accuracy of the proposed and other methods on R.Olmos et.al. dataset.

Methods	Dataset-1		Dataset-2		Dataset-3		Dataset-4	
	Training ratio		Training ratio		Training ratio		Training ratio	
	30%	60%	30%	60%	30%	60%	30%	60%
Fisher's LDA	70%	72%	70%	72%	70%	72%	70%	72%
SVM	87.3%	88%	87.3%	88%	87.3%	88%	87.3%	88%
VGGNet	91.1%	91.5%	87.3%	88%	87.3%	88%	87.3%	88%
ZFNet	92.4%	92.8%	87.3%	88%	87.3%	88%	87.3%	88%
GoogleNet	92.7%	93%	87.3%	88%	87.3%	88%	87.3%	88%
ResNet	92.6%	92.9%	87.3%	88%	87.3%	88%	87.3%	88%
DenseNet	95.8%	96%	87.3%	88%	87.3%	88%	87.3%	88%
PA	97.2%	98.8%	87.3%	88%	87.3%	88%	87.3%	88%
PA + Attention module	97.4%	98.4%	87.3%	88%	87.3%	88%	87.3%	88%

Table 7
Classification accuracy of the proposed and other methods on dataset of D.Ramerio et. al.

Methods	Training ratio	
	50%	80%
Fisher's LDA	70%	72%
SVM	87.3%	88%
VGGNet	91.1%	91.5%
ZFNet	92.4%	92.8%
GoogleNet	92.7%	93%
ResNet	92.6%	92.9%
DenseNet	95.8%	96%
PA	97.2%	98.8%
PA + Attention module	97.4%	98.4%

Table 8
Classification accuracy of the proposed and other methods on ITU firearm dataset.

Methods	Training ratio	
	50%	80%
Fisher's LDA	70%	72%
SVM	87.3%	88%
VGGNet	91.1%	91.5%
ZFNet	92.4%	92.8%
GoogleNet	92.7%	93%
ResNet	92.6%	92.9%
DenseNet	95.8%	96%
PA	97.2%	98.8%
PA + Attention module	97.4%	98.4%

4.3.5. Dataset of D. Ramerio et al.

As discussed earlier the dataset contains huge number of images collected from internet. Property of the dataset is, in this dataset different positions of gun has been considered. They designed the dataset with the images having gun but in different position in images. Performance of the methods are tabulated in Table 7. The result can be summarized as follows:

- GSNet with attention module outperforms in this dataset. State-of-the-art method also achieve competitive result compared to the GSNet.
- GSNet without attention module fails to achieve good accuracy but outperforms baseline methods.
- Baseline methods are failed to distinguish images with gun present in unusual position such as top view of the gun captured.

The results can be expressed as follows:

- Residual attention module and use of the attention module after each dense block restrict the network to loose the shallower lower level features. This preservation of features help the network to distinguish between the images having gun in different position. Hence the backbone network of the GSNet fails here.
- As the baseline networks extract the global semantics, baseline network cannot able to detect the top view of the gun. In the top view of gun global shape of the gun not recognizable.
- Among the baseline methods residual based networks performed competitively as in residual networks also a percentage of shallower features are preserved.

Fig. 7 has shown the classification accuracy of proposed network in tackling challenging conditions. As the dataset consider different views of gun many of the other challenging situation are excluded. As per the Fig. 6 in the dataset, there are type of challenging conditions. It can be noticed that GSNet performed well in each situation. Comparatively performance of the GSNet decreases in distinguishing top viewed guns. Detection of top viewed gun is still a challenge in classification of scenes based on presence of gun.

4.3.6. ITU firearm dataset

Table 8 showing results of the methods on the ITU firearm dataset. The dataset also designed by collecting images from internet. Authors include several challenging conditions to make this benchmark dataset. But the dataset is designed for detection. As per our observation, the dataset contains guns in different position but the guns are covering an enough amount of pixel in a n image. More specifically, guns are clearly visible and cannot be considered as a small object. The results are discussed as follows:

- Baseline methods performed well in this dataset even residual based baseline method outperform the backbone structure of he GSNet without attention module.
- GSNet with attention module performed little better than the other methods.
- Methods performed better under the training ratio of 50% than under the training ratio of 80%

The results can be expressed as follows:

- The baseline methods are distinguish scenes based on the interpretation of global features. In this dataset guns are not considered as small object, guns are represented by a huge number of pixels in the images of this dataset. Hence baseline methods outperforms other methods.
- Proposed architecture without the attention module not able to highlight the global semantics and hence fails miserably.
- Attention modules able to highlight key local semantics and therefore, GSNet with attention module performed little better.
- The dataset contain a large number of images and therefore under 50% training ratio proposed networks and others networks too saturate.

Fig. 7 presents the classification accuracy of GSNet on challenging images of ITU firearm dataset. The results has shown that the GSNet is able to achieve fair accuracy in all challenging scenario. The dataset does not contain significant challenging scenes. The dataset contain images having guns capturing a large number of pixels. Hence, it is easy to interpret scenes based on the guns.

Table 9
Influence of dense-block on different datasets.

Methods	Dense-1		Dense-2		Dense-3		Dense-4	
	Training ratio		Training ratio		Training ratio		Training ratio	
	30%	60%	30%	60%	30%	60%	30%	60%
TUVD-CSA	97.2%	98.8%	87.3%	88%	87.3%	88%	87.3%	88%
IMFDB	97.4%	98.4%	87.3%	88%	87.3%	88%	87.3%	88%
Dataset of Olmos et.al	97.4%	98.4%	87.3%	88%	87.3%	88%	87.3%	88%
Dataset of Ramario et.al	97.4%	98.4%	87.3%	88%	87.3%	88%	87.3%	88%
Gun movies dataset	97.4%	98.4%	87.3%	88%	87.3%	88%	87.3%	88%
ITU firearm dataset	97.4%	98.4%	87.3%	88%	87.3%	88%	87.3%	88%

Table 10
Influence of convolution layer number in each dense block of proposed approach on different datasets.

Methods	Conv-1		Conv-2		Conv-3		Conv-4		Conv-5	
	Training ratio		Training ratio		Training ratio		Training ratio		Training ratio	
	30%	60%	30%	60%	30%	60%	30%	60%	30%	60%
TUVD-CSA	97.2%	98.8%	87.3%	88%	87.3%	88%	87.3%	88%	87.3%	88%
IMFDB	97.4%	98.4%	87.3%	88%	87.3%	88%	87.3%	88%	87.3%	88%
Dataset of Olmos et.al	97.4%	98.4%	87.3%	88%	87.3%	88%	87.3%	88%	87.3%	88%
Dataset of Ramario et.al	97.4%	98.4%	87.3%	88%	87.3%	88%	87.3%	88%	87.3%	88%
Gun movies dataset	97.4%	98.4%	87.3%	88%	87.3%	88%	87.3%	88%	87.3%	88%
ITU firearm dataset	97.4%	98.4%	87.3%	88%	87.3%	88%	87.3%	88%	87.3%	88%

It can be noted that the backbone network of proposed architecture without attention module (SAN module) performed competitively and outperform all the state-of-the-art and baseline methods.

4.3.7. Discussion on other small object dataset

From the results mentioned in the previous section it is highlighted that the proposed method performs well on all the datasets related to firearm to assess the efficacy of the proposed method on other small object dataset, we included UCM dataset [56]. UCM dataset [56] is a aerial image scene classification dataset. Proposed methodology has been implemented on UCM dataset [56] (satellite imagery dataset). UC Merced (UCM) Land Use Dataset is an benchmark dataset containing 21 scene categories and 2100 samples in total. Images in this dataset are all captured via airplane platform. This dataset is characterized by highly similar categories such as dense residential, medium residential and sparse residential to distinguish similar spatial structures with different densities. Proposed methodology implemented on the same dataset with same hyperparameter setting. And it performs good on the same dataset. The accuracy recorded is 97.75%. Whereas MIDC-NET [12] achieved 97% on the same dataset, and from this we can conclude that the proposed algorithm performs competitively with MIDC-Net (algorithm for satellite image classification).

4.4. Performance of proposed methodology in terms of feature extraction

In this work, a deep neural network incorporating an attention module is proposed for scene classification based on the presence of small objects, specifically firearms. The primary objective of the proposed method is to enhance the extraction of small-object features by leveraging deep hierarchical feature representations. To evaluate its effectiveness as a feature extractor, CBAM [42] and BAM [43] attention modules were also implemented on the TUVD-CSA dataset, achieving accuracies of 87% and 89%, respectively. This indicates that the proposed network architecture outperforms existing attention-block-based networks.

Unlike traditional feature fusion techniques, the proposed network refines features hierarchically within Dense Blocks and SAN Modules, ensuring continuous enhancement at multiple stages. To further assess its efficiency, FPN [21] (it explicitly fuse features across multiple scales) was applied to the TUVD-CSA dataset for small-object detection, achieving an accuracy close to 98%. However, instead of explicit multi-scale fusion, the proposed methodology enhances feature propagation through Dense Blocks, where each layer receives inputs from all preceding layers, promoting stronger feature reuse and gradient flow. Additionally, SAN Modules within each Dense Block refine spatial and channel-wise information simultaneously, improving feature selectivity and reducing redundant activations. As a result, the proposed network achieves an accuracy of 98% or higher, demonstrating its superior feature extraction capability.

4.5. Ablation study

4.5.1. Influence of dense block number

Baseline dense structure consist of 4 dense block wherein there are three dense block used in the proposed architecture. To further investigate influence of the fourth dense block, we compare our framework with the original dense structure under both attention module and enhancement layer.

Table 9 lists the related results. We ca observe that:

- Under all the experiments using either without attention module or with attention module performs better while maintaining the fourth dense block performs the worst.
- We also compare said variants of methodologies under both the training ratio 50% and 80%. Under larger training ratio the original network structure with fourth dense block performed similarly as proposed network. But under smaller training ratio, it has a worse performance.

The outcomes can be explained as follows:

- Our proposed network employs less number of convolution layers with attention module to highlight relevant local semantics of the scene. Thus we are avoiding over fitting problem of Deep Networks.
- Proposed architecture with forth dense block perform similarly with proposed network under larger training ratio, which in turn implied that fourth dense block does not have any affect on the performance. Under fewer training samples, use of fourth dense block make the network to over-fit hence it achieves worst accuracy.
- Our proposed structure is more orderly with the removal of the forth dense block. It could be beneficial for multilevel feature representation.

Table 11
Influence of convolution layer number in each dense block of proposed approach on different datasets.

Dataset	PM without enhancement layer		PM with enhancement layer	
	Training ratio		Training ratio	
	30%	60%	30%	60%
TUVD-CSA	97.2%	98.8%	87.3%	88%
IMFDB	97.4%	98.4%	87.3%	88%
Dataset of Olmos et.al	97.4%	98.4%	87.3%	88%
Dataset of Ramario et.al	97.4%	98.4%	87.3%	88%
Gun movies dataset	97.4%	98.4%	87.3%	88%
ITU firearm dataset	97.4%	98.4%	87.3%	88%

Table 12
Influence of augmentation of datasets on the performance of proposed approach.

Dataset	PM implemented on dataset with augmentation	PM implemented on dataset without augmentation
TUVD-CSA	97.2%	83.2%
IMFDB	98.4%	77.3%
Dataset of Olmos et.al	98%	82.5%
Dataset of Ramario et.al	98.4%	90%
Gun movies dataset	98.4%	97.3%
ITU firearm dataset	98.4%	96.4%

4.5.2. Influence of convolution layer number

As mentioned previously, in original dense net structure number of convolution layer (in each dense block) is higher than the number of convolution layers (in each dense block) used in proposed structure. More specifically, number of convolution layer is equal to the number of 3 × 3 convolution layer in each dense block. In original dense structure, dense block contain six(6) 3 × 3 convolution layer wherein our proposed structure contains three(3) 3 × 3 convolution layers. To evaluate the affect of number of 3 × 3 convolution layer on the performance of the proposed structure we report overall accuracy of GSNet under different training ratio when the number of 3 × 3 convolution layer is 1, 2, 3, 4 and 5 respectively. The results are tabulated in Table 10

From the results we can observed that the accuracy of the GSNet is increasing with the number of 3 × 3 layers. The accuracy is at its speak when the number of 3 × 3 layer is 3. Afterwards the performance of proposed architecture slightly decreases with further increase of number of 3 × 3 layers.

The outcome can be expressed by the fact that for small objects in different position, three (3) 3 × 3 layers is appropriate for each dense block. And for four(4) or five(5) 3 × 3 convolution layer the structure is more likely to gt over-fitted and accuracy decreases.

It is worth mentioning that when the number of 3 × 3 layers varies from 2 to 5 the classification performance is competitive when compared with state-of-the-art methods. Moreover, when using one (1) 3 × 3 layer is used, till GSNet performed competitively compared to the state-of-the-art methods. Which further implies that without dense connection too GSNet performed well. It might be explained by the utilization of attention module and enhancement layers. After all, relevant local semantics are highlighted.

4.5.3. Influence of enhancement layer

To demonstrate the influence of the enhancement layer in the proposed network we implement the network structure without the enhancement layer on all the dataset under both the larger and smaller training ratios. All the results are listed in Table 11.

From these results, we can observe that without the enhancement layer results are slightly decreasing compared to the results observed with the enhancement layer.

This outcome could be explained by the following aspects. 1 × 1 convolution layers of enhancement block given higher weights to the relevant semantic label. Use of RELU activation function, the traditional linear feature representation boosted to non-linear representation and the feature representation ability is further enhanced.

The number of convolution channels in this 1 × 1 convolutional layer is equal to the number of inputted feature maps. Hence, each feature map can be refined and the capability to represent scene features is increased with more 565 parameters.

4.5.4. Influence of proposed augmentation procedure

To evaluate the impact of data augmentation, we applied the proposed methodology to both the original and augmented datasets. Results (Table 12) show that augmentation significantly improves testing accuracy for datasets like TUVD-CSA, IMFDB, and R. Olmos et al. which contain varied lighting conditions, complex backgrounds, occlusions, and false positives. In contrast, datasets such as the Gun Movies Database, Knives Images Database, and ITU Firearm Database showed minimal performance changes with augmentation.

To analyze the individual effects of augmentation, we separately applied brightness adjustment (±20%), flipping and rotation, and cropping the gun region. The combined approach yielded better improvements, as shown in Table 12. Notably, cropping had no significant impact except on the TUVD-CSA dataset, likely due to its inclusion of occluded weapon images, whereas other datasets did not account for occlusion during preparation.

4.6. Limitations

Despite GSNet's effectiveness in enhancing small object detection through the use of Small Object Attention (SAN) and enhancement layers, there are still some limitations. The model, while optimized for small object detection, may face challenges in extreme occlusion scenarios or in cases with severe environmental noise, where even the attention mechanism may struggle to capture relevant features. Additionally, the computational efficiency, though improved compared to DenseNet, could be further enhanced for large-scale real-time applications. Future work could focus on refining the attention mechanism to handle more complex scenarios, optimizing the network's architecture for lower latency, and exploring transfer learning techniques to improve performance on diverse datasets without extensive retraining.

5. Conclusion

Summary

In this study, we proposed an enhanced convolutional neural network (CNN) architecture, GSNet, to address the challenges of weapon-based scene classification, specifically focusing on detecting small objects like firearms. The network incorporates DenseNet as a backbone, along with Small Object Attention (SAN) blocks, enhancement layers, and transition layers, which collectively help in preserving low- and mid-level features while emphasizing small, critical objects within a scene. Our model outperformed state-of-the-art methods, achieving a significant accuracy improvement, with a 98% average accuracy across multiple datasets compared to DenseNet's 94.2%. The additional enhancement layers after each transition layer contributed to the model's ability to retain structured and discriminative features, particularly for small object detection in complex environments.

Implications

proposed GSNet model has direct and valuable implications for surveillance and security systems. By improving the accuracy and efficiency of automatic scene classification based on the detection of small weapons, GSNet enhances real-time threat detection in high-stakes environments, such as airports, government facilities, public gatherings, and sensitive border areas. The ability to precisely classify scenes based on the presence of firearms can significantly improve automated surveillance systems, leading to quicker response times and improved public safety.

Recommendations

researchers, we recommend further exploring attention-based mechanisms like the SAN block to enhance feature representation, especially in tasks that involve small object detection. Future work could also focus on optimizing training procedures, experimenting with different learning rate schedules, and using advanced weight initialization techniques such as He initialization for faster convergence. For practitioners, implementing the proposed network architecture into real-world surveillance systems can lead to more accurate and timely threat detection, potentially preventing incidents in high-security environments. Furthermore, utilizing the insights from GSNet's performance could help develop more specialized systems for other critical applications in the security and surveillance domain.

CRediT authorship contribution statement

Rajib Debnath: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kakali Das:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Mrinal Kanti Bhowmik:** Writing – review & editing, Visualization, Validation, Supervision, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work presented here is being conducted in the Computer Vision Laboratory of Computer Science and Engineering Department, Tripura University (A Central University), Suryamaninagar-799022, Tripura (W), India. The work is supported by DST-SERB International Research Experience (SIRE), India Fellowship awarded to Mrinal Kanti Bhowmik for the year 2022–2023 under the Grant No. SIR/2022/000387, Dated: 12/05/2022 from Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India.

Data availability

Data will be made available on request.

References

[1] A. Karp, Estimating Global Civilian-held Firearms Numbers, Briefing Paper, Small Arms Survey, Small Arms Survey, Department of Foreign Affairs and Trade, Australia, 2018, URL https://books.google.co.in/books?id=NjNwuwEACAAJ.

[2] Desmond U. Patton, William R. Frey, Michael Gaskell, Guns on social media: Complex interpretations of gun images posted by Chicago youth, Palgrave Commun. 5 (119) (2019) 1–8.

[3] Deng Jia, Dong Wei, Socher Richard, Li Li-Jia, Li Kai, Fei-Fei Li, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, 2009, pp. 248–255.

[4] Roberto Olmos, Siham Tabik, Francisco Herrera, Automatic handgun detection alarm in videos using deep learning, Neurocomputing 275 (2018) 66–72.

[5] Sydney Maples Justin Lai, Developing a Real-Time Gun Detection Classifier, World academy of science, Stanford University, 2017.

[6] T. Radhika, V. Chandrasekar, Analysis of Markovian jump stochastic Cohen–Grossberg BAM neural networks with time delays for exponential input-to-state stability, Neural Process. Lett. 55 (2023) 11055–11072.

[7] Yang Cao, A. Chandrasekar, T. Radhika, V. Vijayakumar, Input-to-state stability of stochastic Markovian jump genetic regulatory networks, Math. Comput. Simulation 222 (2024) 174–187.

[8] Gui Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, Xiaoqiang Lu, AID: A benchmark data set for performance evaluation of aerial scene classification, IEEE Trans. Geosci. Remote Sens. 55 (7) (2017) 3965–3981.

[9] S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions, Int. J. Uncertain. Fuzziness Knowl. Based Syst. 6 (2) (2011) 107–116.

[10] L. Zhang, Y. Han, Y. Yang, M. Song, S. Yan, Q. Tian, Discovering discriminative graphlets for aerial image categories recognition, IEEE Trans. Image Process. 22 (12) (2013) 5071–5084.

[11] Junwei Han, Peicheng Zhou, Dingwen Zhang, Gong Cheng, Lei Guo, Zhenbao Liu, Shuhui Bu, Jun Wu, Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding, ISPRS J. Photogramm. Remote. Sens. 89 (2014) 37–48.

[12] Qi Bi, Kun Qin Zhili Li, Han Zhang Kai Xu, Gui-Song Xia, A multiple-instance densely-connected ConvNet for aerial scene classification, IEEE Trans. Image Process. 29 (2020) 4911–4926.

[13] Yuanfei Huang Yanting Hu, Xinbo Gao, Channel-Wise and spatial feature modulation network for single image super-resolution, IEEE Trans. Circuits Syst. Video Technol. 30 (11) (2020) 3911–3927.

[14] Yu Chen, Hongbing Meng, Xinling Wen, Pengge Ma, Yuxin Qin, Zhengxiang Ma, Zhaoyu Liu, Classification methods of a small sample target object in the sky based on the higher layer visualizing feature and transfer learning deep networks, EURASIP J. Wirel. Commun. Netw. 127 (2018).

[15] Paul Tresson, Dominique Carval, Philippe Tixier, William Puech, Hierarchical classification of very small objects: Application to the detection of arthropod species, IEEE Access 9 (2021) 63925–63932.

[16] Jiahe Zhang, Xiongkuo Min, Jun Jia, Zehao Zhu, Jia Wang, Guangtao Zhai, Fine localization and distortion resistant detection of multi-class barcode in complex environments, Multimedia Tools Appl. (2020) 1–20.

[17] Jun Jia, Guangtao Zhai, Jiahe Zhang, Zhongpai Gao, Zehao Zhu, Xiongkuo Min, Xiaokang Yang, Guodong Guo, EMBDN: An efficient multiclass barcode detection network for complicated environments, IEEE Internet Things J. 6 (6) (2019) 9919–9933.

[18] Jun Jia, Guangtao Zhai, Ping Ren, Jiahe Zhang, Zhongpai Gao, Xiongkuo Min, Xiaokang Yang, Tiny-BDN: An efficient and compact barcode detection network, IEEE J. Sel. Top. Signal Process. 14 (4) (2020) 688–699.

[19] Adnan Sharif, Guangtao Zhai, Jun Jia, Xiongkuo Min, Xiangyang Zhu, Jiahe Zhang, An accurate and efficient 1-D barcode detector for medium of deployment in IoT systems, IEEE Internet Things J. 8 (2) (2021) 889–900.

[20] Adnan Sharif, Guangtao Zhai, Xiongkuo Min, Jun Jia, Kashif Munir, Enhancing decoding rate of barcode decoders in complex scenes for IoT systems, IEEE Internet Things J. 8 (24) (2021) 17495–17507.

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.

[22] Xiangyang Liu, Yujia Wang, Shunping Yu, Zhilu Zhang, AFPN: Asymptotic feature pyramid network for object detection, Neurocomputing 450 (2021) 217–226.

[23] Mingxing Tan, Ruoming Pang, Quoc V. Le, EfficientDet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10781–10790.

[24] Yang Zhu, Yufei Su, Xinyu Li, Jingdong Wang, Chunhua Lin, CM-Net: Concentric mask-based arbitrary-shaped text detection, IEEE Trans. Image Process. 29 (2020) 5261–5273.

[25] Minghui Zhang, Jian Wang, Xin Liu, Yi Liu, Lei Zhang, Zoom text detector: A zoom-in mechanism for text detection in natural scene images, Pattern Recognit. 95 (2019) 216–225.

[26] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, Jiaya Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8759–8768.
[27] Minghui Liao, Yue Zou, Zhenbo Wan, Pengyuan Lyu, Cong Yao, Xiang Bai, Reinforcement shrink-mask for arbitrary-shaped text detection, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2022) 2435–2447.
[28] Samet Akcay, Mikolaj E. Kundegorski, Chris G. Willcocks, Toby P. Breckon, Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery, IEEE Trans. Inf. Forensics Secur. 13 (9) (2018) 2203–2215.
[29] Mai Kamal el den Mohamed, Ahmed Taha, Hala H. Zayed, Automatic gun detection approach for video surveillance, Int. J. Sociotechnol. Knowl. Dev. (IJSKD) 12 (1) (2020) 49–66.
[30] David Romero, Christian Salamea, Design and proposal of a database for firearms detection, in: The International Conference on Advances in Emerging Trends and Technologies, 2019, pp. 348–360.
[31] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, Xinping Guan, A multimodal saliency model for videos with high audio-visual correspondence, IEEE Trans. Image Process. 29 (2020) 3805–3819.
[32] Xiongkuo Min, Guangtao Zhai, Ke Gu, Xiaokang Yang, Fixation prediction through multimodal analysis, ACM Trans. Multimed. Comput. Commun. Appl. 13 (1) (2016) 1–23.
[33] Tahereh Hassanzadeh, Daryl Essam, Ruhul Sarker, Evolutionary attention network for medical image segmentation, in: 2020 Digital Image Computing: Techniques and Applications, DICTA, 2020, pp. 1–8.
[34] Yangyang Li, Qin Huang, Xuan Pei, Yanqiao Chen, Licheng Jiao, Ronghua Shang, Cross-layer attention network for small object detection in remote sensing imagery, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14 (2021) 2148–2161.
[35] Zhelin Li, Lining Zhao, Xu Han, Mingyang Pan, Lightweight ship detection methods based on YOLOv3 and DenseNet, Math. Probl. Eng. 2020 (1) (2020) 4813183.
[36] Shudi Wang, Manman Xu, Ying Sun, Guozhang Jiang, Yaoqing Weng, Xin Liu, Guojun Zhao, Hanwen Fan, Jun Li, Cejing Zou, et al., Improved single shot detection using DenseNet for tiny target detection, Concurr. Comput.: Pr. Exp. 35 (2) (2023) e7491.
[37] Minghao Zhai, Junchen Liu, Wei Zhang, Chen Liu, Wei Li, Yi Cao, Multi-scale feature fusion single shot object detector based on densenet, in: Intelligent Robotics and Applications: 12th International Conference, ICIRA 2019, Shenyang, China, August 8–11, 2019, Proceedings, Part V 12, 2019, pp. 450–460.
[38] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700–4708.
[39] Zequn Qin, Pengyi Zhang, Fei Wu, Xi Li, FcaNet: Frequency channel attention networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 783–792.
[40] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, Kyoung Mu Lee, Channel attention is all you need for video frame interpolation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, pp. 10663–10671, (07).
[41] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[42] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.
[43] Jongchan Park, Sanghyun Woo, Joon-Young Lee, In-So Kweon, BAM: Bottleneck attention module, 2018.
[44] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2021) 3349–3364.
[45] IMFDB, The Internet Movie Firearms Database. http://www.imfdb.org/wiki/main_page.
[46] Andrzej Matiolański Michał Grega, Piotr Guzik, Mikołaj Leszczuk, Automated detection of firearms and knives in a CCTV image, Sensors 16 (1) (2016) 47.
[47] Javed Iqbal, Muhammad Akhtar Munir, Arif Mahmood, Afsheen Rafaqat Ali, Mohsen Ali, Orientation aware object detection with application to firearms, 2019, arXiv preprint arXiv:1904.10032.
[48] Rajib Debnath, Mrinal Kanti Bhowmik, A novel framework for automatic localization of gun carrying by moving person using various indoor outdoor mimic and real time views/scenes, IET Image Process. 14 (17) (2021) 4663–4675.
[49] Guangtao Zhai, Xiongkuo Min, Perceptual image quality assessment: a survey, Sci. China Inf. Sci. 63 (211301) (2020).
[50] Xiongkuo Min, Ke Gu, Guangtao Zhai, Xiaokang Yangand Wenjun Zhang, Patrick Le Callet, Chang Wen Chen, Screen content quality assessment: Overview, benchmark, and beyond, ACM Comput. Surv. 54 (9) (2021) 1–36.
[51] Xiongkuo Min, Ke Gu, Guangtao Zhai, Jing Liu, Xiaokang Yang, Chang Wen Chen, Blind quality assessment based on pseudo-reference image, IEEE Trans. Multimed. 20 (8) (2018) 2049–2062.
[52] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, Xiaokang Yang, Blind image quality estimation via distortion aggravation, IEEE Trans. Broadcast. 64 (2) (2018) 508–517.
[53] Xiongkuo Min, Kede Ma, Ke Gu, Guangtao Zhai, Zhou Wang, Weisi Lin, Unified blind quality assessment of compressed natural, graphic, and screen content images, IEEE Trans. Image Process. 26 (11) (2017) 5462–5474.
[54] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Mylène C.Q. Farias, Alan Conrad Bovik, Study of subjective and objective quality assessment of audio-visual signals, IEEE Trans. Image Process. 29 (2020) 6054–6068.
[55] Rajib Debnath, Mrinal Kanti Bhowmik, A comprehensive survey on computer vision based concepts, methodologies, analysis and applications for automatic gun/ knife detection, J. Vis. Commun. Image Represent. 78 (2021) 103165.
[56] Yi Yang, Shawn Newsam, Geographic image retrieval using local invariant features, IEEE Trans. Geosci. Remote Sens. 51 (2) (2013) 818–832.

Rajib Debnath received his Bachelor of Technology (Information Technology) degree from Bengal Institute of Technology and Management, Shantiniketan, West Bengal, India, in 2010. Masters in Technology (Computer Science and Engineering) and Doctor of Philosophy (Ph.D) under Computer Science and Engineering Department of Tripura University (A Central University), Suryamaninagar, Tripura, India in 2012 and 2022 respectively. Currently working as an Assistant Professor in Computer Science and Engineering Department, GITAM (Deemed to be University), Hyderabad Campus, Hyderabad, Telangana, 502329, India. His topics of interest are related to the field of Computer Vision, Object Detection, Machine Learning, Image and Video Processing etc.

Kakali Das received her Bachelor of Engineering (Computer Science and Engineering) degree from Tripura Institute of Technology, Tripura, India in 2012, Masters in Technology (Computer Science and Engineering) degree from Tripura University (A Central University), Suryamaninagar, Tripura, India in 2014, and Doctor of Philosophy (Ph.D) degree under Computer Science and Engineering Department of Tripura University (A Central University), Suryamaninagar, Tripura, India. Currently working as an Assistant Professor in Computer Science and Engineering Department, GITAM (Deemed to be University), Hyderabad Campus, Hyderabad, Telangana, 502329, India. Her topics of interest are related to the field of Medical Image Processing, Infrared Imaging, Machine Learning, Computer Vision etc.

Mrinal Kanti Bhowmik received the B.E. degree in computer science and engineering from the Tripura Engineering College in 2004, the M.Tech. degree in computer science and engineering from Tripura University (A Central University), India, in 2007, and the Ph.D. degree in engineering from Jadavpur University, Kolkata, India, in 2014. He also spent Fall 2022 as a DST-SERB International Research Experience Scholar with SIRE Fellowship, Government of India at the NYU Center for Cybersecurity (CCS), Tandon School of Engineering, New York University, New York City. He has successfully completed two Department of Electronics and Information Technology (DeitY) (Now Ministry of Electronics and Information Technology (MeitY)) funded projects, one the Department of Biotechnology (DBT)-Twinning project, one Society for Applied Microwave Electronics Engineering and Research (SAMEER) funded project, one Indian Council of Medical Research (ICMR), and one Defense Research and Development Organization (DRDO), Government of India funded project. He is currently the Principal Investigator of a DBT, Government of India funded project and Co-Principal Investigator of ICMR, Government of India funded project. Since July 2010, he has served with the Department of Computer Science and Engineering, Tripura University as an Assistant Professor and from March, 2023 he has been serving with Department of Computer Science and Engineering, Tripura University as an Associate Professor. He was awarded the Short Term Indian Council of Medical Research (ICMR), Department of Health Research (DHR) International Fellowship from 2019 to 2020 as a Senior Indian Biomedical Scientist for bilateral cooperation in cross-disciplinary research area (i.e., biomedical diagnostic and inferencing systems). He has also published a sole authored book in the domain of computer vision, published by CRC Press, Taylor and Francis Group, Chapman and Hall Book. His current research interests are in the field of machine learning, computer vision, security and surveillance, medical imaging, and biometrics.