APSIPA Transactions on Signal and Information Processing, 2025, 14, e21

Original Paper

Scene Understanding by Fused Hu's Invariant Moments and Deep Learning Features

Michael Nachipyangu* and Jiangbin Zheng

Northwestern Polytechnical University, China

ABSTRACT

Convolutional neural networks (CNN) are widely used in the recognition and classification of scene images due to their effectiveness in this task. However, their applicability is not quite as favorable when used with variations of parameters such as rotation, scaling, and translation in input data. To overcome this drawback, this study presents a feature fusion technique that combines Hu moments with deep learning features derived from the CNN model. Hu's moments of an image are statistical values obtained based on the intensities of the image pixels that are invariant to geometric transformations. These moments are then combined with the features of the fully connected layer of the CNN model, making the proposed method more accurate and robust. The study also utilizes data augmentation, specifically geometrical transformations such as rotating, scaling, flipping, and translation to balance class image distribution in training datasets and reduce interclass bias resulting from the imbalance in number of images within different classes. The fused feature representation was evaluated on three benchmark datasets: MIT67, AID and Scene15. Detailed experiments with different CNN models were conducted, and Inception-ResNetV2 as deep feature extractor combined with Hu Moments demonstrated the effectiveness of the proposed approach which delivers significant improvements in accuracy scores, Scene15: 96.4%,

*Corresponding author: michael.nachipyangu@mail.nwpu.edu.cn.

Received 11 February 2025; revised 19 June 2025; accepted 15 July 2025
ISSN 2048-7703; DOI 10.1561/116.20250007

AID: 94.1% and MIT67: 87.1%. This result presents a novel avenue approach for enhancing the resilience and accuracy of Scene Understanding.

Keywords: CNN, data augmentation, deep learning features, Hu's moments, scene understanding

1 Introduction

Scene understanding is a foundational challenge in computer vision,requiring the extraction and interpretation of meaningful information from visual data to enable machines to recognize, categorize, and reason about complex environments, including human-computer interaction, robotic movement, and autonomous systems [43]. It is the process of obtaining clear patterns of visuals with the ability to perform various tasks, including categorization of scenes, identification of objects, and navigation of space. Traditionally, this field utilized handmade features, which were engineered patterns of extrinsic characteristics that identified the representation of particular image properties, including edges, texture, or color distribution [8]. However, these methods formed the basis for pragmatically solving the problems of visual scene analysis, but had significant drawbacks associated with generalization abilities and the handling of nonlinear transformations.

The advent of deep learning has significantly advanced the field of scene understanding by enabling the automatic extraction of hierarchical features from raw image data. Deep neural networks, particularly convolutional neural networks (CNN), have demonstrated remarkable capability in recognition accuracy and made CNNs the key component of most image-based categorization methods [21, 16, 34, 23, 24, 29, 41, 20, 38]. Despite these successes, CNN can sometimes be sensitive to geometric transformations such as translation, scaling, and rotation, which are common in real-world applications. Several techniques have been proposed in the literature to improve generalization, including data augmentation. Where the training data set is created artificially using techniques such as flipping, scaling, rotation, or translation, among others, similar to the explanation above, data augmentation not only extends the variety of samples used for training, but also acts as a way to reduce the risk of overfitting [22]. As noted in [30], data augmentation, when used in conjunction with transfer learning, has been shown to help overcome the problem of limited training samples and improve model performance. These transformations require that specific features, which should remain unchanged, be preserved when the scene image undergoes geometrical changes for image recognition. This requirement extends beyond simple data augmentation to

realistic application areas, such as robotic applications and self-driving cars, where images may be transformed in unintended ways due to environmental factors.

Hu [11] proposed a seven-moment invariant that remains unchanged despite the geometric transformation including rotation, scaling, and translation. These features, computed from the pixel intensity distribution of images, have been found to be useful in attaining geometric invariance. The study of [17] investigated the nature and behavior of such moment invariants under different image resolutions and observed that as the resolution increases, it enhances the stability of moment invariants. We used this study to determine the optimal image size for our experiments. In addition, the work done in [4]introduced Hu's moment invariants to a binary problem, where there were only two classes taken from the Scene15 dataset: the MIT-street and MIT-highway classes. They achieved an improved classification rate, despite using only a limited number of classes for the experiments. This study presents a novel method that integrates deep learning features with Hu's moment invariants to improve scene understanding. Convolutional Neural Networks are employed to extract a substantial number of salient features that contain semantic and structural information, which is crucial for discrimination tasks. These features are then combined with Hu's moment invariants, yielding a representation that preserves information and remains invariant to geometric transformations. By combining these two complementary feature sets, the proposed method addresses limitations found in traditional CNN-based models, offering enhanced stability and classification accuracy for scene understanding applications. The framework's performance is evaluated on three benchmark datasets: MIT67, AID, and Scene15. Different scene categories in these datasets are challenging and diverse enough to assess the performance of the proposed approach. Sample images from MIT67 are shown in Figure 1. The experiments demonstrate that integrating deep features and Hu's moments enhances classification performance, highlighting the significance of this approach for scene understanding in real-world contexts. This integration enhances the scene recognition aspect of the model, resulting in more consistent recognition even when scenes are rotated, scaled, or translated in some manner. In addition, data augmentation strategies are incorporated into the method to enhance its ability to generalize.

The primary contribution of this paper is the development of a new hybrid feature architecture that combines deep learning features with Hu moments for scene understanding. The remaining part of this paper is structured as follows. Section 2 reviews related work, including several recent deep learning techniques for scene understanding, and the use of Hu moments in image-based tasks. Section 3 explains the method that has been proposed and gives details of how deep features are merged with Hu moments, along with the overall approach. Section 4 describes the experiment, the data to be used,

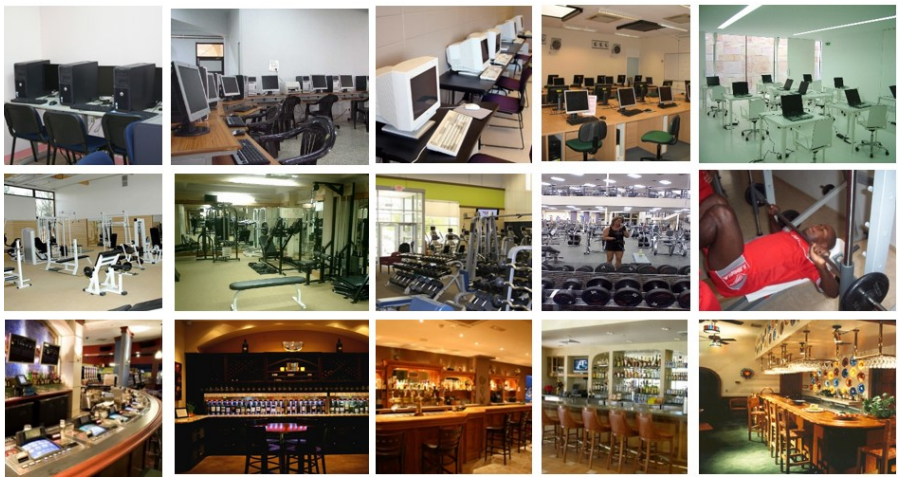

Figure 1: Sample images of scene categories from MIT67 dataset. First row: Computer room; second row: Gym; third row: Bar.

data pre-processing, and data measures used to assess the results, and it also gives results and discussions on the application of the idea. Section 5 shows the conclusion and part of future work in which the effectiveness of the introduced method is described and a comparison with similar approaches is made.

2 Related Work

2.1 Deep Learning Approaches for Scene Understanding

In recent years, there has been a growing interest in using deep learning methods to understand scenes [21, 41, 10, 39, 13, 12]. These methods have achieved remarkable success in various computer vision applications, including image recognition, object detection, and semantic segmentation. However, most existing methods rely solely on deep features extracted from a single modality, disregarding the potential benefits of incorporating multimodal information. In [21], one of the remarked approaches employs the ResNet architecture and the multi-layer fusion strategy to retain discriminative features across different layers. This strategy solves one of the main issues with transfer learning, namely the fact that higher-layer features may obscure fine details in the lower layers. The authors only used CNN which to some extent lack geometric transformation invarience. To combat intraclass variability in indoor scene classification, [13] proposed a new feature transformation method, which focuses especially on mid-level features. It is evident that elements

give an acceptable level of abstraction on one side and sufficient detail on the other which makes them useful when considering the classification performance. The method tried to get the features that balance between the high and low levels image presentation but still didn't look at the features which remain unaltered when subjected to changes. The work of [12] extended scene classification by introducing multi-stage feature fusion in InceptionResNetV2. This method combines the set of local features and the set of global features for the stages and gets the fused classifier to yield a better accuracy. Likewise, [33] discussed the effects of repeated transformations in CNN outputs on detail loss. To enhance the efficiency of the discovered models and the overall model resilience, they suggested that GoogLeNet should be divided into three parts and use the classifier product rule at each part. The work of [37] addresses the challenge of limited labeled samples in remote sensing image classification by proposing a discrete wavelet-based multi-level deep feature fusion method. The method involves multiple steps including discrete wavelet transform, deep feature extration, and a modified discriminant correlation analysis. The method is computationally expensive but also depend on careful selection of parameters for each step so as to have better results at the end. There have been some more work which specifically try to tackle a certain challenge like [28] focusses on reducing computational cost for scene classification of UAV images, the author use Modified GhostNet model to improve the challenges of distinguishing ground objects from UAV images. The similar idea from [2] where the author address the problem f remote sensing scene classification with limited labelled data. They proposed the method called RS-FewShotSSL which used Deep learning model that has to be trained using high-resolution and low-resolution images. Aforementioned studies use different techniques trying to get more discriminative features by leveraging features at the fully connected layer, multi-layer fusion, hybrid feature fusion, stage-wise integration, or mid-level features, but still these are features only obtained by the CNN model extractors which are still suffer the problem on not being invariant to geometric transformations and some still has performance which need to be improved.

2.2 Application of Hu's Moment in Image-based Classification

Moment Invariants, particularly those developed by Hu, have been widely applied in various image processing tasks and methods for their ability to handle rotation, scaling, and translation of images. The authors in [18] showed that when Hu's Moment Invariants were integrated into a method for detecting automotive connectors on manufacturing assembly lines it offered a picture of 5.03% higher matching accuracy and 45.06% improved speed as compared to other standard techniques. To address the computational workload and the issues related to precision of classical methods, the study employed a gain

function and an adaptive pyramid search approach. Furthermore, [27] implemented Hu's Moment Invariants for the early diagnosis of cervical cancer with a feature extraction of cervical cell images and classification with SVM. Their method means that the accuracy rate was 71.9% with the processors taking 0.98705 seconds, thus underlining the use of CNNs together with Hu's Moment Invar-iants can be used for much higher levels of accuracy in classification in much less time. Additionally, similar to [35], the use of Hu's Moments in identifying patterns for medical images was confirmed to help in making medical images invariant to the scales and rotations. Furthermore, the authors stressed issues related to the experimental inputs and directives in addition to segmentations necessary while applying Hu's Moments for feature extraction in medical image processing. By drawing on the concept of Hu, [4] has been able to show the role of moment Invariants by doing experiments on two classes of Scene15 (MIT-street, and MIT-Highways) datasets and claimed to have improved the classification accuracy. This study was more like a binary classification problem as the author only categorize two classes only. However, in an attempt to enhance classification and recognition of binary objects in image processing, computer vision and machine learning applications, [14] integrated Local Binary Patterns with Hu's Moments. Through this combination, the model's handling of the transformed appearance of objects was further improved. The author in [25] address the problem of geometric transformation when classifying ships in video surveillance images. They incorporated denoising, segmentation and Hu's Moment into their CNN framework and created a compact vector of Hu's Mo-ments; including this compact vector with deep learning features gave them a better discriminant image representation. The authors proved that the presented approach based on the suggested method performs better than state-of-art methodologies in the context of accuracy and complexity. This study set a new avenue where the deep learning features and Hu's moment are used for scene understanding.

3 Proposed Method

This research proposes the improvement of CNN architectures by incorporating Hu's Moment Invariants into the feature extraction segment to improve image transformation invariance. Seven invariant Moments of Hu were obtained from these images and we combined them with the fully connected layer of our CNN model as shown in Figure 2. For feature extraction, the CNN model as well as the Hu's Moment sub-pipeline was run in parallel. The images undergo preprocessing which involved resizing of the images to a dimension of 224 by 224 pixels. Regarding the deep learning model, the images were preserved in the shape of the three-dimensional color channels.

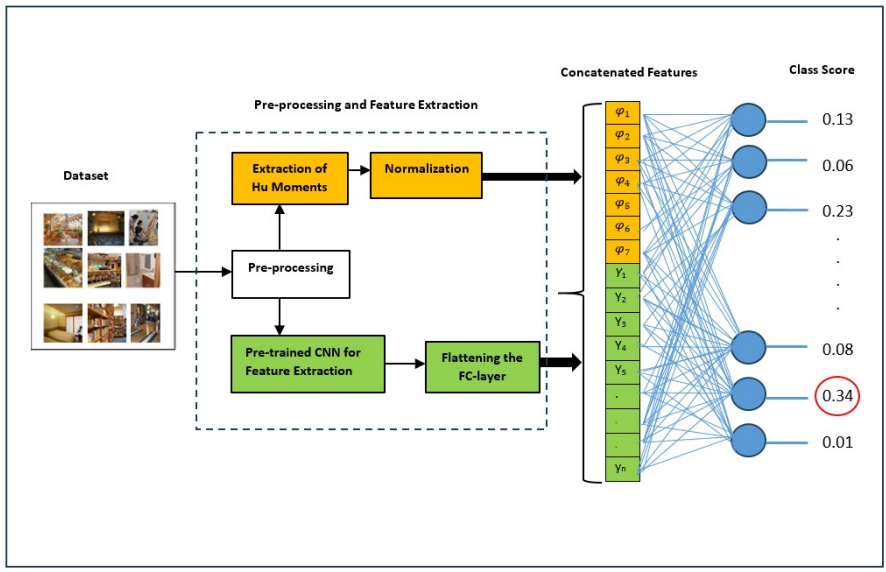

Figure 2: The proposed method architecture.

Before the images were fed into the Hu's Moment extractor, they were shifted to gray scale. The fully connected layer features that were extracted were concatenated with Hu's Moments that were obtained from the Hu's Moment extraction module. All these features were concatenated for training the classifier.

3.0.1 Deep Learning Feature Extraction

Features for deep learning were extracted with three CNN models that were pre-trained from ImageNet dataset[3] which include: ResNet50 [9], Inception-ResNetV2 [31], and InceptionV3 [32].

These models have shown top performances in most computer vision tasks. Due to the high computational costs and time considerations, transfer learning was adopted to save of training them from scratch. From the above models, we formulated the feature set that incorporates discriminative features of CNN that were incorporated with the Hu moment-invariant features to form a new appended feature vector.

3.1 Hu's Invariant Moment

The Hu Moments Invariant features are derived as defined by the works [11, 17, 1]. The two-dimensional $(p+q)^{th}$ order moments are defined as follows:

$$m_{pq} = \int \int_{-\infty}^{\infty} x^p y^q f(x,y) dxdy \tag{1}$$

where $p, q = 0, 1, 2, ...$

The invariant features can be achieved by using central moments which can be defined as follows;

$$\mu_{pq} = \int \int_{-\infty}^{\infty} (x-\bar{x})^p (y-\bar{y})^q f(x,y) dxdy \tag{2}$$

where $p, q = 0, 1, 2..., \bar{x} = \frac{m_{10}}{m_{00}}$ and $\bar{y} = \frac{m_{10}}{m_{00}}$ Given the Centroid of the image $f(x,y)$ with the pixel point at $\bar{x}, \bar{y}$, the centroid moments μ_{pq} whose center has been shifted to the centroid of the image. This makes the central moments invariant to image translation. The Scale Invariance can be attained by normalization of the central moments [30] The normalized central moments are defined as follows;

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \gamma = \frac{p+q+2}{2}, p+q = 2, 3, ... \tag{3}$$

Based on the normalized central moments, [11] introduced seven moments invariants.

$\phi_1 = \eta_{20} + \eta_{02}$
$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$
$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \mu_{03})^2$
$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \mu_{03})^2$
$\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + 3\eta_{02})^2]$
$+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \mu_{03})^2]$
$\phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$
$\phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$
$- (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$

These seven moment invariants have unchanged properties when the image is subjected to scaling, translation and rotation. By using these seven moment invariants feature vector can be expressed as;

$$\nu = (\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \phi_7)$$

The obtained feature vector V is normalized to ensure that all features have a similar scale, preventing some features from dominating others due to their large magnitudes and making the learning algorithm converge faster [7].

3.2 Features Fusion

In order to create a more comprehensive image representation, the high-level features extracted from Deep learning models were combined with the Hu Moments features of each image to form a composite vector. The process involved the applying cascade fusion, which directly connects the feature maps while retaining all elements. This fusion technique was proposed by [25]. Concretely, it is especially well suited to fuse feature maps of different dimensions.

3.3 Classification

Our proposal builds upon the combined Deep learning and Hu invariant moments features after feature concatenation followed by a fully connected layer for classification. Instead, fused features allow the classification to take advantage of the deep learning and the Hu moments features. The task included several classes; the classifier became Softmax, and the cost function became the modified categorical-cross entropy. The output of the last layer was fixed at 15 for the Scene15 [15] datasets, 30 for the AID [36] datasets and 67 for the MIT-67 [24] dataset, respectively.

4 Experiments and Results Discussion

4.1 Experimental Setup

The proposed approach was implemented by Python 3.10 using PyCharm IDE on 64-bit Windows 11 running on a Dell OptiPlex 7000 operated with a 12th Gen Intel(R) Core (TM) i7-12700T CPU @ 1.70GHz and installed RAM of 64.00 GB. The experiment was conducted to validate our proposed approach. In implementation, three datasets were used, which were MIT67 [24], AID [36] and Scene15 [15].

4.2 Datasets

The some datasets used were observed to have an imbalanced image number across different classes. The imbalance can affect the calculation of some metrics; thus, the Data Augmentation was applied to balance the number of images distributions within the classes. The graphs in Figure 3 show the number of image distribution between classes of the benchmark datasets used to validate our proposed approach. We addressed the imbalanced by using Data Augmentation; technique used involved rotation, zooming, translation, and flipping. Sample images of augmented images are shown in Figure 4. This technique balanced the number of images but also increased the image variations hence boost generalizability of our approach. The datasets were

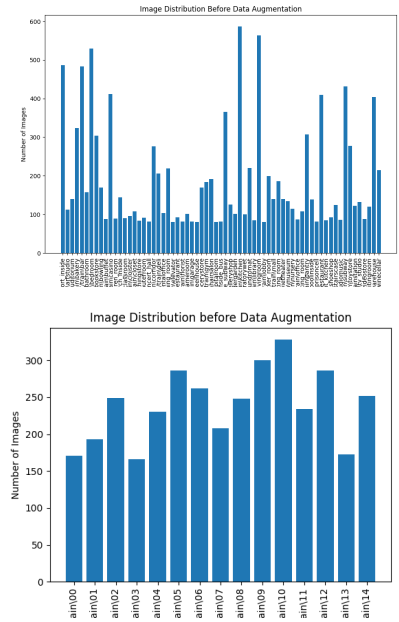

Figure 3: The image distribution in MIT67 and Scene15.

divided in ratios of 80%, 10%, and 10% for training, test, and validation sets respectively. To avoid any chance of the system to see the training set, these three sets were stored in different folders.

4.3 Evaluation Metrics

Confusion matrix generate values which are extremely valuable used to obtain important metrics for machine learning models. These metrics are Accuracy, Precision and Recall. Accuracy is the total number of correct classifications divided by the total number of classification [6], given by the following formula: TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is the ratio between the number of images correctly classified as TP and the total number of images in the class under observation (TP and FP) [6]. Precision can be defined as the accuracy of the classification of a particular class.

$$P = \frac{TP}{TP + FP}$$

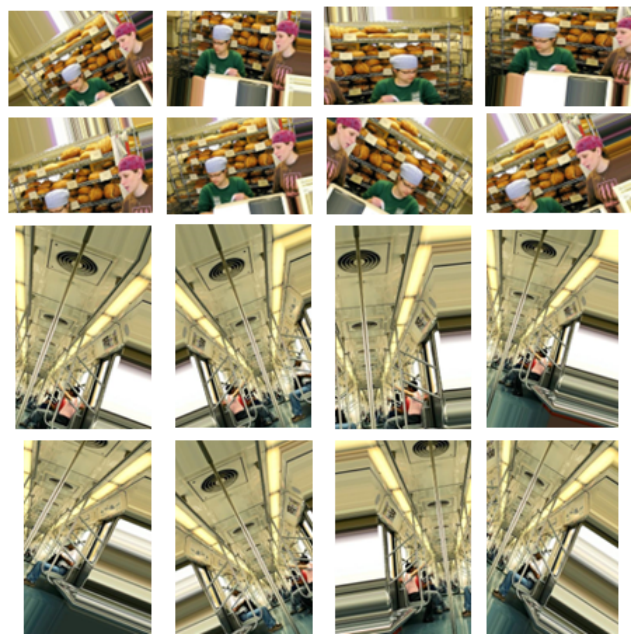

Figure 4: Data augmented samples from MIT67 dataset; above is Bakery and the lower is Inside Subway categories.

Recall is the ratio between the total number of images classified as TP and the total number of images in the class under observation [6].

$$R = \frac{TP}{TP + FN}$$

These metrics were used to measure the effectiveness of the models used in the experiments on both datasets.

4.4 Ablation Study

4.4.1 Data Augmentation

The contribution of data augmentation was demonstrated through an experimental study in which geometric data augmentation techniques, specifically rotation, translation, scaling, and shearing were employed to simulate various real-world scenarios. Three distinct model architectures were evaluated, and results were recorded both with and without the application of data augmentation. The findings indicate a consistent improvement in performance in all models when data augmentation was utilized.

4.4.2 Feature Importance

To quantify the contribution of Hu's moments, we conducted an ablation study comparing three configurations: (1) CNN alone (InceptionResNetV2), (2) Hu's moments alone and (3) the proposed fusion method. Table 1 reports results on MIT67, AID and Scene. 87. 1%, outperforming the CNN baseline (84%) and Hu's Moments alone (67. 3%) for MIT67, 94. 1%, outperforming CNN alone (93%) and Hu's Moments alone (73. 2%) for AID, and 96. 4%, outperforming CNN alone (95%) and Hu's Moments alone (75%) for Scene15, this confirms that the combination enhances robustness We also did feature importance analysis by using SHAP (SHapley Additive exPlanations) where the more important features that contributed to the prediction appeared more on the fusion setup, as shown in Figure 5.

Table 1: Effects of data augmentation on different models for Scene15, AID and MIT67.

Model	Augmentation	Accuracy (%)			Precision (%)			Recall (%)		
		Scene15	AID	MIT67	Scene15	AID	MIT67	Scene15	AID	MIT67
ResNet50	No	86	84.6	46	84.9	85	48	85	85	48
InceptionV3	No	85	84	60	86.1	84.5	61.5	86	84.5	61.5
InceptionResNetV2	No	94	91.5	81	94.5	92	82	94	92	82
ResNet50	Yes	88	86	65	87	87	64	87	86.9	64
InceptionV3	Yes	92	88	72	93	89	73	92	89	72
InceptionResNetV2	Yes	95	93	84	96	93.8	85	96	94.1	85

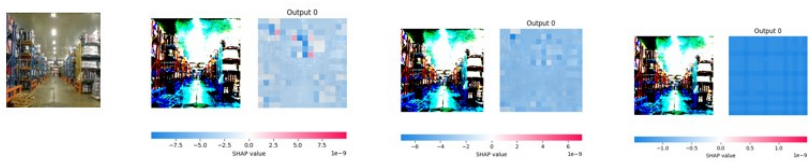

Figure 5: From left: warehouse image, SHAP values for fused feature, CNN alone and Hu's moment alone.

4.5 Results Discussion

This research demonstrates that integrating Hu's moments, deep learning features, and data augmentation significantly improves the performance of image classification. The study evaluated the effectiveness of fusing geometric invariants with learned features using three pre-trained ImageNet CNN models - ResNet50, InceptionResNetV2, and InceptionV3. The results indicate that models utilizing the combined features outperformed those relying solely on the DL features, showing improvements in accuracy, robustness, and generalization. This can be seen in Tables 2, 3 and 4 where the results have

Table 2: Models' perfomance on Scene15 with and without Hu's moment.

Without Hu's Moment				With Hu's Moment		
Model	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
InceptionResNetV2	95	96	96	96.4	97	97
ResNet50	88	87	87	92.3	92	92
InceptionV3	92	93	92	93	94	94

Table 3: Models' performance on AID with and without Hu's moment.

Without Hu's Moment				With Hu's Moment		
Model	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
InceptionResNetV2	93	93.8	94	94.1	95	95
ResNet50	86	87	86.9	87.5	88.2	88
InceptionV3	88	89	89	90	91.2	91

Table 4: Models' performance on MIT67 with and without Hu's moment.

Without Hu's Moment				With Hu's Moment		
Model	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
InceptionResNetV2	84	85	85	87.1	88	88
ResNet50	65	64	64	66.5	66	67
InceptionV3	72	73	72	74	75	75

been tabulated showing the comparison of the results with and without Hu moments.

The integrated approach balanced the use of Hu's moments, which are rotationally, scale, and translation-invariant features, with the discriminative DL features extracted from pre-trained models. This fusion addressed common drawbacks associated with standalone CNNs, including vulnerability to spatial transformations of the input data. The incorporation of Hu's moments enabled the models to capture the precise geometric attributes of the input images, leading to more accurate classification. Moreover, the application of data augmentation enhanced the observed increases in precision by providing a wider set of training data, reducing the likelihood of over-fitting. Techniques such as moving, mirroring, and resizing boosted other aspects of the model by generalizing the future data, particularly in cases of intraclass variance. The combination of data augmentation and feature fusion resulted in greater model robustness due to complementary processes. The image distribution between classes was imbalanced as seen in Figure 3, Data Augmentation was also used to eliminate bias between classes due to the unbalanced image distribution. When Hu's moments were introduced, the InceptionResNetV2 model exhibited the highest increases compared to other CNN models, likely due to its outstanding ability to learn both low-level and high-level features. Similar enhancements were observed for ResNet50 and InceptionV3, demonstrating the efficacy of the proposed approach across various architectures. The analysis of the confusion matrix and quantitative results indicated that the fused models had higher precision, recall, and accuracy compared to the comparative models as tabulated in Table 5.

Table 5: Comparison with other works.

Author	Method	Accuracy (%)		
		Scene15	AID	MIT67
Tang *et al.* [33]	(G-MS2F)	93.37	-	80.3
Zhou *et al.* [40]	PLACES-CNN	92.15	-	79.49
Liu *et al.* [21]	FTOTLM	94.01	-	74.63
Rezanejad *et al.* [26]	Pre-trained CNN using scene contours	-	-	48.61
Guo *et al.* [5]	LS-DHM (VggNet11)	-	-	83.75
Salman *et al.* [13]	DUCA	94.5	-	71.8
Zhou *et al.* [42]	AlexNet + ImageNet	84.23	-	56.79
Liu *et al.* [19]	Deep Learning + Spatial coding	89.70	-	62.90
Nachipyangu *et al.* [22]	Pre-trained CNN+Data Augmentation	95	-	86
Devendran *et al.* [4]	Moment Invariant+ANN	-	-	83.5
Xia *et al.* [36]	VGG-VD-16	-	89.64	-
Alosaimi *et al.* [2]	RS-FewShotSSL	-	87.13	-
Shen *et al.* [28]	Modified GhostNet	-	92.05	-
Song *et al.* [37]	DWMLFF	-	86.17	-
Ours	Fused DL and Hu's moment	**96.4**	**94.1**	**87.1**

The study suggests that using Hu's moments in conjunction with DL features provides a more comprehensive feature set, satisfying both geometric invariance and semantic aspects of the images. Additionally, the study found that the proposed approach has the potential to enhance the performance of Scene understanding tasks by addressing the limitations of standalone CNN models. However, the computational complexity associated with extracting Hu's moments and integrating DL features remains a significant limitation, which requires further investigation to optimize the process.

5 Conclusion

This research further demonstrates the potential of integrating Hu's moments, combined with deep learning features and data augmentation, to significantly enhance Scene image understanding performance. The incorporation of geometric invariants and learned features introduced here has addressed important limitations associated with the original CNN models, including their susceptibility to spatial transformations such as rotation, scaling, or translation. The results showcase consistent improvements across three pre-trained CNN models, including ResNet50, InceptionResNetV2, and InceptionV3, underscoring the reliability and broad applicability of the proposed ap-proach. Given that Hu's moments are invariant to image transformations and hence rotation, and the DL features are semantically rich, the fused models have outperformed models relying solely on DL features in terms of accuracy, precision, and recall. Furthermore, the application of data augmentation has provided an even stronger foundation to this framework, increasing variabil-

ity and reducing overfitting, thereby enhancing overall performance. The findings suggest that integrating geometric invariants as additional features into deep learning architectures can expand the range of features covered. Although the high computational overhead remains a limitation, the present study identifies avenues for further research to minimize demands and refine the fusion process, ultimately improving the performance of more comprehensive computer vision applications. Consequently, the proposed approach offers a promising path towards developing more effective and stable Scene Understanding methods that are invariant to illumination and other conditions. This research underscores the significant potential of combining geometric invariants, deep learning features, and data augmentation to achieve substantial improvements in Scene understanding, paving the way for more robust and versatile computer vision solutions.

References

[1] S. AbuRass, A. Huneiti, and M. B. Al-Zoubi, "Enhancing convolutional neural network using Hus moments", *International Journal of Advanced Computer Science and Applications*, 11(12), 2020.

[2] N. Alosaimi, H. Alhichri, Y. Bazi, B. Ben Youssef, and N. Alajlan, "Self-supervised learning for remote sensing scene classification under the few shot scenario", *Scientific Reports*, 13(1), 2023, 433.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database", in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, 248–55.

[4] V. Devendran, H. Thiagarajan, and A. Santra, "Scene categorization using invariant moments and neural networks", in *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, Vol. 1, IEEE, 2007, 164–8.

[5] S. Guo, W. Huang, L. Wang, and Y. Qiao, "Locally supervised deep hybrid model for scene recognition", *IEEE transactions on image processing*, 26(2), 2016, 808–20.

[6] G. Hackeling, *Mastering Machine Learning with scikit-learn*, Packt Publishing Ltd, 2017.

[7] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer, 2009.

[8] M. Hayat, S. H. Khan, M. Bennamoun, and S. An, "A spatial layout and scale invariant feature representation for indoor scene classification", *IEEE Transactions on Image Processing*, 25(10), 2016, 4829–41.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–8.

[10] L. Herranz, S. Jiang, and X. Li, "Scene recognition with cnns: objects, scales and dataset bias", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 571–9.

[11] M.-K. Hu, "Visual pattern recognition by moment invariants", *IRE transactions on information theory*, 8(2), 1962, 179–87.

[12] A. Khan, A. Chefranov, and H. Demirel, "Building discriminative features of scene recognition using multi-stages of inception-ResNet-v2", *Applied Intelligence*, 53(15), 2023, 18431–49.

[13] S. H. Khan, M. Hayat, M. Bennamoun, R. Togneri, and F. A. Sohel, "A discriminative representation of convolutional features for indoor scene recognition", *IEEE Transactions on Image Processing*, 25(7), 2016, 3372–83.

[14] R. Kumar and K. Mali, "Local Binary Pattern for Binary Object Classification using Coordination Number (CN)* and Hu's Moments", in *2021 9th international conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO)*, IEEE, 2021, 1–7.

[15] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Vol. 2, IEEE, 2006, 2169–78.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *nature*, 521(7553), 2015, 436–44.

[17] D. Li, "Analysis of moment invariants on image scaling and rotation", in *Innovations in Computing Sciences and Software Engineering*, Springer, 2010, 415–9.

[18] Y. Li, J. Lu, Y. Song, Z. Zhang, and K. Liu, "Automobile Connectors Recognition Algorithm Based on Improved Hu Invariant Moments", in *2022 International Conference on Informatics, Networking and Computing (ICINC)*, IEEE, 2022, 178–82.

[19] B. Liu, J. Liu, and H. Lu, "Learning representative and discriminative image representation by deep appearance and spatial coding", *Computer Vision and Image Understanding*, 136, 2015, 23–31.

[20] S. Liu, G. Tian, *et al.*, "An indoor scene classification method for service robot Based on CNN feature", *Journal of Robotics*, 2019, 2019.

[21] S. Liu, G. Tian, and Y. Xu, "A novel scene classification model combining ResNet based transfer learning and data augmentation with a filter", *Neurocomputing*, 338, 2019, 191–206.

[22] M. Nachipyangu, J. Zheng, and P. Mawagali, "Transfer Learning and On-Fly Data Augmentation for Scene UnderstandingUsing Inception-ResNet", in *2023 6th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, 2023, 496–500.

[23] G. Patterson, C. Xu, H. Su, and J. Hays, "The sun attribute database: Beyond categories for deeper scene understanding", *International Journal of Computer Vision*, 108, 2014, 59–81.

[24] A. Quattoni and A. Torralba, "Recognizing indoor scenes", in *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, 413–20.

[25] Y. Ren, J. Yang, Q. Zhang, and Z. Guo, "Ship recognition based on Hu invariant moments and convolutional neural network for video surveillance", *Multimedia Tools and Applications*, 80, 2021, 1343–73.

[26] M. Rezanejad, G. Downs, J. Wilder, D. B. Walther, A. Jepson, S. Dickinson, and K. Siddiqi, "Scene categorization from contours: Medial axis based salience measures", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, 4116–24.

[27] B. P. Sari and Y. Jusman, "Classification system for cervical cell images based on hu moment invariants methods and support vector machine", in *2021 International Conference on Intelligent Technologies (CONIT)*, IEEE, 2021, 1–5.

[28] X. Shen, H. Wang, B. Wei, and J. Cao, "Real-time scene classification of unmanned aerial vehicles remote sensing image based on Modified GhostNet", *PloS one*, 18(6), 2023, e0286873.

[29] J. Shi, H. Zhu, S. Yu, W. Wu, and H. Shi, "Scene categorization model using deep visually sensitive features", *IEEE Access*, 7, 2019, 45230–9.

[30] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning", *Journal of big data*, 6(1), 2019, 1–48.

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning", in *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31, No. 1, 2017.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 2818–26.

[33] P. Tang, H. Wang, and S. Kwong, "G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition", *Neurocomputing*, 225, 2017, 188–97.

[34] R. Tombe and S. Viriri, "Remote sensing image scene classification: Advances and open challenges", *Geomatics*, 3(1), 2023, 137–55.

[35] S. Urooj and S. P. Singh, "Geometric invariant feature extraction of medical images using Hu's invariants", in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, 2016, 1560–2.

[36] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, 55(7), 2017, 3965–81.

[37] S. Yang, H. Wang, H. Gao, and L. Zhang, "Few-shot remote sensing scene classification based on multi subband deep feature fusion", *Math. Biosciences Eng*, 20(7), 2023, 12889–907.

[38] D. Zeng, M. Liao, M. Tavakolian, Y. Guo, B. Zhou, D. Hu, M. Pietikäinen, and L. Liu, "Deep learning for scene classification: A survey", *arXiv preprint arXiv:2101.10531*, 2021.

[39] H. Zeng, X. Song, G. Chen, and S. Jiang, "Learning scene attribute for scene recognition", *IEEE Transactions on Multimedia*, 22(6), 2019, 1519–30.

[40] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, "Places: An image database for deep scene understanding", *arXiv preprint arXiv:1610.02055*, 2016.

[41] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition", *IEEE transactions on pattern analysis and machine intelligence*, 40(6), 2017, 1452–64.

[42] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database", *Advances in neural information processing systems*, 27, 2014.

[43] B. Zhu, J. Xie, X. Gao, and G. Xu, "A fusiform network of indoor scene classification with the stylized semantic description for service-robot applications", *Expert Systems with Applications*, 245, 2024, 122979.