www.ijerst.com
Vol. 22, No. 1(1), 2026

Research Paper

CIFAKE: Explainable Image Classification and AI-Generated Synthetic Image Identification

[1]M NARESH, [2]Dr.P SURI BABU

[1]PG Scholar, Dept. of CSE, KMM Institute of Technology & Science, Ramireddipalle, Tirupati, AP, India.

[2]Professor, Dept. of CSE, KMM Institute of Technology & Science, Ramireddipalle, Tirupati, AP, India.

Abstract: Recent advancements in synthetic data synthesis have made it harder for humans to discern between real-world and AI-generated images. This work addresses the critical need for data reliability and authenticity by providing a method to enhance the identification of AI-generated images via computer vision algorithms. The study uses a synthetic dataset built using latent diffusion and modeled after the CIFAR-10 dataset to build a set of images for comparison with real photographs. The classification challenge is given as a binary issue that involves distinguishing between real and artificial intelligence-generated pictures. A CNN2D model, which contains 32 neurons and yields the best results, is employed for this. Two convolutional layers, two MaxPooling2D layers, and two dense layers make up this structure. Additionally, the Grad-CAM (Gradient Class Activation Mapping) technique highlights components that aid CNN in distinguishing between real and fake images. A modified version that was optimized without the inclusion of additional layers like dropout or global average pooling achieved 95.94% accuracy compared to 94.98% for the proposed CNN2D. The performance is evaluated using accuracy, precision, recall, F1-score, and a confusion matrix.

Index terms - *ai-generated images, synthetic image detection, cnn2d, grad-cam, image classification, latent diffusion, explainable ai, cifar-10, deep learning, binary classification, fake image detection, computer vision, performance evaluation, neural networks*

1. INTRODUCTION

Recent years have seen a rapid advancement in the field of artificial intelligence (AI)-generated synthetic pictures, and recognizing AI-generated images is now crucial to ensuring the accuracy of image data. In contrast to the recent past, when generative technologies routinely produced images with severe visual faults that were evident to the human eye, we today face the possibility of AI models creating high-fidelity and photorealistic images in a matter of seconds. AI-generated visuals are now capable of competing with humans and winning art competitions due to their increasing quality. Latent Diffusion Model (LDM) generative models have emerged as a powerful method for producing synthetic images.

These recent developments have fundamentally altered our ideas of truth, authenticity, and creativity. As a result, consumer-level technology that is easily exploited for fraud and privacy infractions is now available. These philosophical and sociological implications raise significant questions about the nature of reality and dependability at the forefront of contemporary technology. Thanks to recent technological advancements, it is now feasible to produce images of such high quality that individuals cannot tell the difference between a real-life snapshot and one that is only a hallucination of the weights and biases of an artificial neural network.

2. LITERATURE SURVEY

2.1 High-Resolution Image Synthesis with Latent Diffusion Models:

ABSTRACT: Diffusion models (DMs) divide the image production process into a sequential application of denoising autoencoders to achieve state-of-the-art synthesis results on picture data and beyond. Additionally, its formulation enables a guiding mechanism to control the picture-generating process without retraining. However, inference is expensive due to sequential assessments because these models often operate directly in pixel space, and optimizing powerful DMs sometimes requires hundreds of GPU days. To enable DM training on limited processing resources while preserving their flexibility and quality, we employ them in the latent space of powerful pretrained autoencoders. Unlike previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between lowering complexity and keeping detail, greatly improving visual fidelity. By including cross-attention layers into the model architecture, we turn diffusion models into robust and flexible generators for wide conditioning inputs like as text or bounding boxes. Convolutional high-resolution synthesis is made possible by this. Our latent diffusion models (LDMs) achieve highly competitive performance on a range of tasks, including unconditional image generation, text-to-image synthesis, and super-resolution, as well as new state-of-the-art scores for image inpainting and class-conditional image synthesis, with much lower computational requirements than pixel-based DMs.

2.2 Predicting image credibility in fake news over social media using multi-modal approach:

ABSTRACT: Fake images are mostly spread on social media. Pictures altered with software or other methods to change their meaning are called fakes. Misinformation and polarization disseminated via tweeting result from false images. Identity verification on social media is essential to combat fake photos. Text is often connected to phony images. Therefore, a multi-modal architecture with textual and visual feature learning is implemented. Few multi-modal frameworks exist, and those that do require extra tasks to determine modalities' associations. This study presents an efficient multi-modal strategy for spotting fake microblogging platform photos. No more subcomponents needed. Suggested design uses sentence transformer for text processing and explicit convolution neural network model EfficientNetB0 for photos. Combining dense layers of visual and linguistic feature embeddings predicts fake photos. Testing the model on MediaEval (Twitter) and Weibo yields accuracy forecasts of 85.3% and 81.2%, respectively. We evaluate the model using the latest Twitter dataset, which contains photographs of India's important

events in 2020. Experimental results show that the proposed model outperforms cutting-edge multi-modal frameworks.

2.3 On the use of Benford's law to detect GAN-generated images:

ABSTRACT: Thanks to Generative Adversarial Network (GAN) designs, anybody may make lifelike synthetic images. False news, opinion formation, and other social and political consequences may come from malevolent GAN-generated images. Therefore, detection-capable technologies are needed to restrict the widespread spread of false images. This study examines if Benford's rule can distinguish GAN-generated photos from real ones. Benford's law describes quantized DCT coefficients' most significant digit distribution. We show that expanding and generalizing this characteristic may extract a compact feature vector from an image. This feature vector may be fed to a simple classifier to recognize GAN-generated pictures.

2.4 Zero-Shot Text-to-Image Generation:

ABSTRACT: The primary objective of text-to-image generation has always been to identify better modeling assumptions for training on a particular dataset. These assumptions may include complex structures, additional losses, or accidental information during training, such as segmentation masks or object component labels. For this, we describe a simple approach based on a transformer that autoregressively expresses the text and picture tokens as a single stream of data. When there is sufficient data and scalability, our approach is competitive with previous domain-specific models when evaluated in a zero-shot manner.

2.5 Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding:

ABSTRACT: We present Imagen, a text-to-image diffusion model with unmatched photorealism and a deep level of language understanding. Imagen uses the power of huge transformer language models for text understanding and the strength of diffusion models for producing high-fidelity images. Our main finding is that generic large language models (such as T5) are surprisingly good at encoding text for image synthesis when pretrained on text-only corpora: expanding the language model in Imagen significantly improves sample fidelity and image-text alignment more than expanding the image diffusion model. Imagen achieves a new state-of-the-art FID score of 7.27 on the COCO dataset without ever training on COCO, and human raters deemed Imagen samples to be equivalent to the COCO data itself in image-text alignment. To assess them more completely, we provide DrawBench, a comprehensive and rigorous benchmark for text-to-image models. We compare Imagen with more recent methods like VQ-GAN+CLIP, Latent Diffusion Models, and DALL-E 2 using DrawBench. In side-by-side comparisons, we find that human raters prefer Imagen over other models in terms of sample quality and image-text alignment.

3. METHODOLOGY

i) Proposed Work:

By incorporating new deep learning layers such Global Average Pooling and Dropout, the expanded suggested system improves the original CNN2D model to more successfully categorize actual and AI-generated pictures. These improvements concentrate on strengthening feature extraction and enhancing the

model's capacity for generalization. The Global Average Pooling layer ensures that the model focuses only on the most important characteristics by reducing dimensionality while maintaining important spatial information. In addition, the Dropout layer encourages the model to acquire more resilient and varied representations by randomly deactivating neurons during training, hence preventing overfitting.

Grad-CAM is used to visually emphasize significant picture areas that affect the model's predictions, increasing confidence in AI judgments and making the system more comprehensible and user-friendly. Furthermore, a Flask-based web application is developed that enables users to input photos and obtain real-time categorization results with visual heatmaps. In addition to ensuring that the model operates with more accuracy and dependability, this thorough and comprehensible methodology makes it feasible for real-world deployment by enabling users to comprehend the reasons behind each choice.

ii) System Architecture:

The system architecture is created as a simplified pipeline that combines explainability, categorization, and image processing into a seamless flow. The first step is input image acquisition, where users use a Flask-based web interface to contribute either actual or AI-generated photographs. The Modified CNN2D model, which has two convolutional layers, two MaxPooling layers, Global Average Pooling, and Dropout layers, receives these preprocessed pictures. To identify whether an image is synthetic or real, the model conducts binary classification after extracting deep characteristics from the picture.

After classification, heatmaps that show the precise areas of the picture that affected the model's choice are produced using the Grad-CAM (Gradient-weighted Class Activation Mapping) approach. These heatmaps, which provide visual explanations to improve trust and transparency, are shown in the user interface alongside the categorization result. In order to evaluate and improve the model, the system also has performance assessment modules that calculate accuracy, precision, recall, F1-score, and confusion matrix values. High accuracy and end-user usability are guaranteed by this modular and interpretable design.

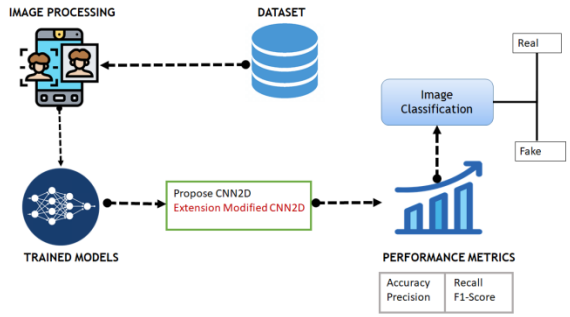

Fig.1. Proposed Architecture

iii) MODULES:

- **Data loading:** We will import the dataset using this module.
- **Image Processing:** In order to improve the quality of photos and extract useful information, image processing entails modifying and evaluating them. Images must be prepared for efficient categorisation and analysis using methods including resizing, shuffling, normalising, and dividing datasets into training and testing groups.
- **Model generation:** Model construction: CNN2D proposal and extension CNN2D was

altered. Each algorithm's performance assessment metrics are computed.

- **Admin login:** This module allows the administrator to log in.
- **Drug Side Effect Prediction:** This module allows users to upload test data.
- **Prediction:** The final prediction was shown.

iv) ALGORITHMS:

CNN2D: A convolutional neural network designed for image classification applications, CNN2D automatically extracts information from images using a number of convolutional and pooling layers. This study uses CNN2D to distinguish between images created by artificial intelligence (AI) and those that are not. By analyzing the collection of images, the model is able to identify visual patterns that differentiate the two groups. This algorithm, which classifies the images with high accuracy and consistently identifies them, is the fundamental model.

Modified CNN2D: The Modified CNN2D algorithm enhances the original CNN2D design by including additional layers such as Global Average Pooling and Dropout layers. This modification aims to improve feature extraction and reduce overfitting by focusing on the most crucial traits and removing the less crucial ones. By analyzing the same dataset of real and artificial intelligence-generated images, the project's improved model offers increased robustness and classification accuracy. The modifications ensure that the model is well optimized for identifying subtle differences across image classes.

4. EXPERIMENTAL RESULTS

The proposed system was evaluated using a combination of real images from the CIFAR-10 dataset and synthetic images generated using latent diffusion models. The CNN2D and Modified CNN2D architectures were trained and tested on this binary classification task to distinguish between real and AI-generated images. The Modified CNN2D model outperformed the base version, achieving an improved accuracy of 95.94%, compared to 94.98% by the original CNN2D. Key performance metrics such as precision, recall, and F1-score also demonstrated high effectiveness in correctly classifying both classes, confirming the robustness of the model.

To validate model interpretability, Grad-CAM was used to produce heatmaps that visually identified the regions in each image that contributed most to the classification decision. This helped in verifying that the model was focusing on relevant features and not background noise. The confusion matrix further showed a low rate of misclassification, reinforcing the model's precision. Overall, the experimental results confirmed that the integration of feature-optimizing layers and explainability tools led to a highly accurate, interpretable, and efficient system for detecting AI-generated synthetic images.

Accuracy: A test's accuracy is determined by its capacity to distinguish between healthy and ill cases. To gauge the accuracy of the test, find the percentage of examined instances that had true positives and true negatives. According to the computations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: Precision is the number of affirmative cases or the classification's accuracy rate. The following formula is applied to assess accuracy:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: A model's ability to recognise every instance of a pertinent machine learning class is measured by its recall. The ratio of accurately predicted positive observations to the total number of positives indicates how well a model can identify class instances.

$$Recall = \frac{TP}{(FN + TP)}$$

mAP: Mean Average Precision is one ranking quality metric (MAP). It considers the number of relevant recommendations and their position on the list. MAP at K is calculated as the arithmetic mean of the Average Precision (AP) at K for each user or query.

$$\boldsymbol{mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k}$$

$$\boldsymbol{AP_k = \text{ the AP of class } k}$$

$$\boldsymbol{n = \text{ the number of classes}}$$

F1-Score: An accurate machine learning model is indicated by a high F1 score. combining precision and recall to increase model correctness. The accuracy statistic quantifies the frequency with which a model correctly predicts a dataset.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

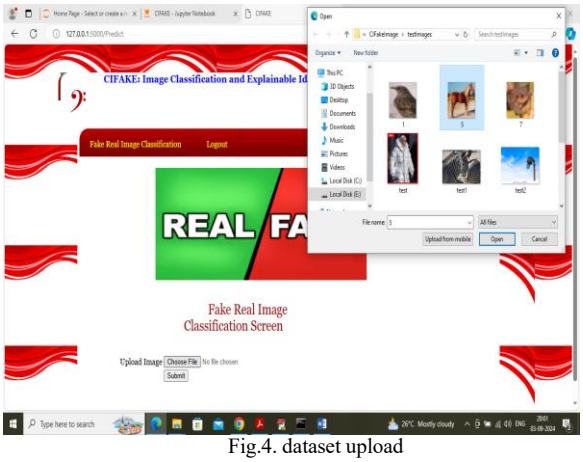

Fig.4. dataset upload

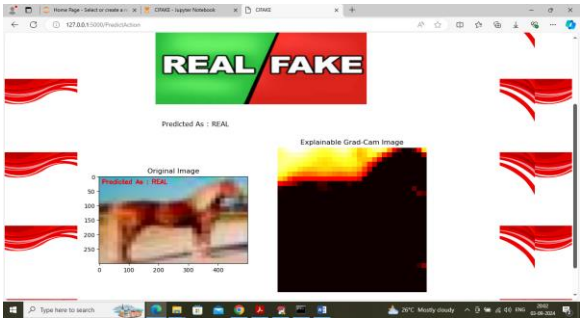

Fig.5. results

	ML Model	Accuracy	Precision	Recall	f1_score
0	Propose CNN2D	94.98	95.02	94.98	94.98
1	Extension- Modified CNN2D	95.94	95.99	95.94	95.94

Fig.6.accuracy table results

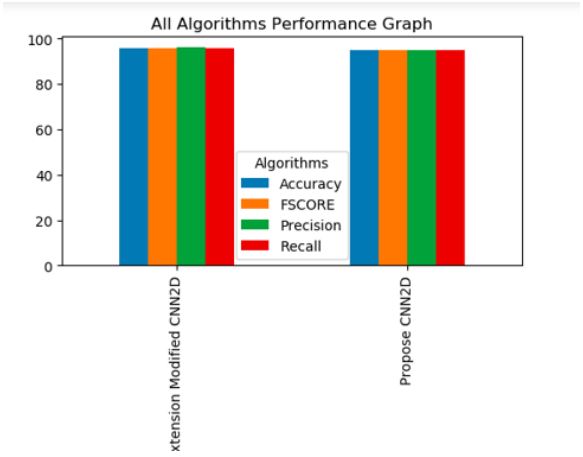

Fig.6. graphical representation

5. CONCLUSION

In summary, the work successfully addresses the challenge of distinguishing between real and artificial intelligence-generated images, a task that has become more challenging as synthetic data generation has advanced. Using a powerful CNN2D algorithm, the system achieves an outstanding accuracy rate of 94.98%, demonstrating its ability to classify photographs according to their validity. The Modified CNN2D method, which attains an incredible accuracy of 95.94%, also yields even better results. These results show the effectiveness of convolutional neural networks in image classification tasks and support the approach of employing both original and modified architectures to increase performance. By providing insights into the network's decision-making process, explainability techniques like Grad-CAM improve the model and help users understand which characteristics affect categorization outcomes. All things considered, the project not only achieves high classification accuracy but also improves the field of computer vision and image processing research by offering insightful data and practical methods for the identification of AI-generated content in subsequent studies.

6. FUTURE SCOPE

Future research may look at attention-based approaches for dataset classification to enhance the model's performance and offer more methods for explainable AI. Adding images from other fields, such human faces and clinical scans, and updating the dataset with advanced synthetic imagery might broaden the scope of this study. This modification will not only improve classification accuracy but also make the proposed technique more relevant across a broader variety of areas, ultimately leading to a deeper understanding of AI-generated content detection.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, ‘‘High-resolution image synthesis with latent diffusion models,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 10684–10695.

[2] B. Singh and D. K. Sharma, ‘‘Predicting image credibility in fake news over social media using multi-modal approach,’’ Neural Comput. Appl., vol. 34, no. 24, pp. 21503–21517, Dec. 2022.

[3] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro, ‘‘On the use of Benford’s law to detect GAN-generated images,’’ in Proc. 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, pp. 5495–5502.

[4] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, ‘‘Zero-shot text-to-image generation,’’ in Proc. Int. Conf. Mach. Learn., 2021, pp. 8821–8831.

[5] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, ‘‘Photorealistic text-to-image diffusion models with deep language understanding,’’ 2022, arXiv:2205.11487.

[6] D. Deb, J. Zhang, and A. K. Jain, ‘‘AdvFaces: Adversarial face synthesis,’’ in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), Sep. 2020, pp. 1–10.

[7] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, ‘‘Model inversion attack: Analysis under gray-box scenario on deep learning based face recognition system,’’ KSII Trans. Internet Inf. Syst., vol. 15, no. 3, pp. 1100–1118, Mar. 2021.

[8] J. J. Bird, A. Naser, and A. Lotfi, ‘‘Writer-independent signature verification; evaluation of robotic and generative adversarial attacks,’’ Inf. Sci., vol. 633, pp. 170–181, Jul. 2023.

[9] G. Pennycook and D. G. Rand, ‘‘The psychology of fake news,’’ Trends Cogn. Sci., vol. 25, no. 5, pp. 388–402, May 2021.

[10] K. Roose, ‘‘An AI-generated picture won an art prize. Artists aren’t happy,’’ New York Times, vol. 2, p. 2022, Sep. 2022.

[11] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, ‘‘Adapting pretrained vision-language foundational models to medical imaging domains,’’ 2022, arXiv:2210.04133.

[12] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, ‘‘Moûsai: Text-to-music generation with long-context latent diffusion,’’ 2023, arXiv:2301.11757.

[13] F. Schneider, ‘‘ArchiSound: Audio generation with diffusion,’’ M.S. thesis, ETH Zurich, Zürich, Switzerland, 2023.

[14] D. Yi, C. Guo, and T. Bai, ‘‘Exploring painting synthesis with diffusion models,’’ in Proc. IEEE 1st Int. Conf. Digit. Twins Parallel Intell. (DTPI), Jul. 2021, pp. 332–335.

[15] C. Guo, Y. Dou, T. Bai, X. Dai, C. Wang, and Y. Wen, ‘‘ArtVerse: A paradigm for parallel human–machine collaborative painting creation in Metaverses,’’ IEEE Trans. Syst., Man, Cybern., Syst., vol. 53, no. 4, pp. 2200–2208, Apr. 2023.

[16] Z. Sha, Z. Li, N. Yu, and Y. Zhang, ‘‘DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models,’’ 2022, arXiv:2210.06998.

[17] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, ‘‘On the detection of synthetic images generated by diffusion models,’’ 2022, arXiv:2211.00680.

[18] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, ‘‘Deepfake video detection through optical flow based CNN,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 1205–1207.

[19] D. Güera and E. J. Delp, ‘‘Deepfake video detection using recurrent neural networks,’’ in Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS), Nov. 2018, pp. 1–6.

[20] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, ‘‘M2TR: Multi-modal multi-scale transformers for Deepfake detection,’’ in Proc. Int. Conf. Multimedia Retr., Jun. 2022, pp. 615–623.

[21] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, ‘‘A hybrid CNNLSTM model for video Deepfake detection by leveraging optical flow features,’’ in Proc.

Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2022, pp. 1–7.

[22] H. Li, B. Li, S. Tan, and J. Huang, ‘‘Identification of deep network generated images using disparities in color components,’’ Signal Process., vol. 174, Sep. 2020, Art. no. 107616.

[23] S. J. Nightingale, K. A. Wade, and D. G. Watson, ‘‘Can people identify original and manipulated photos of real-world scenes?’’ Cognit. Res., Princ. Implications, vol. 2, no. 1, pp. 1–21, Dec. 2017.

[24] A. Krizhevsky and G. Hinton, ‘‘Learning multiple layers of features from tiny images,’’ 2009.

[25] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, ‘‘LAION-5B: An open large-scale dataset for training next generation image-text models,’’ 2022, arXiv:2210.08402.

[26] Y. LeCun, Y. Bengio, and G. Hinton, ‘‘Deep learning,’’ Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[27] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, ‘‘Recent advances in convolutional neural networks,’’ Pattern Recognit., vol. 77, pp. 354–377, May 2018.

[28] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, ‘‘A survey of convolutional neural networks: Analysis, applications, and prospects,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[29] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, ‘‘XAI—Explainable artificial intelligence,’’ Sci. Robot., vol. 4, no. 37, Dec. 2019, Art. no. eaay7120.

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ‘‘Grad-CAM: Visual explanations from deep networks via gradient-based localization,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 618–626.