International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 2, 2026

Peer Reviewed, Referred & Indexed Journal

HYBRID CNN–VISION TRANSFORMER FRAMEWORK FOR ENHANCED LUNG NODULE DETECTION

BANNELA RAJASHEKAR[1]
rajashekar55645@gmail.com
[1]B. Tech Student, Department of Computer Science and Engineering and Business Systems, Rajeev Gandhi Memorial College of Engineering & Technology, Nandyal, Andhra Pradesh

PEDDAMALLU GANGA PRANEETH REDDY[2],
gangapraneethreddy@gmail.com
[2]B. Tech Student, Department of Computer Science and Engineering and Business Systems, Rajeev Gandhi Memorial College of Engineering & Technology, Nandyal, Andhra Pradesh

KARNATI MADHUSUDHAN REDDY[3]
kmsreddy8639@gmail.com
[3]B. Tech Student, Department of Computer Science and Engineering and Business Systems, Rajeev Gandhi Memorial College of Engineering & Technology, Nandyal, Andhra Pradesh

P. ARUN BABU[4],
arunbabu1208@gmail.com
[4].Associate professor, Department of Computer Science and Engineering and Business Systems, Rajeev Gandhi Memorial College of Engineering & Technology, Nandyal, Andhra Pradesh, Nandyal, Andhra Pradesh

ABSTRACT

Lung cancer is one of the leading causes of cancer-related deaths worldwide, and early detection of pulmonary nodules in computed tomography (CT) scans plays a crucial role in improving patient survival. Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have shown significant success in medical image analysis by effectively extracting local spatial features. However, CNN-based models often struggle to capture long-range contextual relationships present in complex CT images, which may limit their detection performance. This study proposes a Hybrid CNN–Vision Transformer (CNN–ViT) framework for enhanced lung nodule detection in CT scans. The proposed approach combines the strengths of CNNs and Vision Transformers to improve feature learning. A 3D Convolutional Neural Network (CNN) is used to extract local texture and structural features from CT scan patches, while the Vision Transformer (ViT) utilizes multi-head self-attention mechanisms to capture global contextual relationships across lung regions. By integrating both local and global feature representations, the hybrid model improves the accuracy and robustness of lung nodule detection. The proposed framework is evaluated using the LUNA16 (Lung Nodule Analysis 2016) dataset derived from the LIDC-IDRI lung CT scan database. Performance evaluation is carried out using standard metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). Experimental results demonstrate that the hybrid CNN–ViT model achieves improved detection performance and better generalization compared to conventional CNN-based methods, making it a promising approach for computer-aided lung cancer diagnosis.

Keywords: Lung Nodule Detection, Convolutional Neural Network (CNN), Vision Transformer (ViT), Medical Image Analysis, CT Scan, Deep Learning, LUNA16 Dataset, Multi-Head Self Attention.

I. INTRODUCTION

Lung cancer is one of the most common and deadly diseases worldwide, accounting for a significant number of cancer-related deaths every year. Early detection of lung cancer is crucial for improving treatment outcomes and increasing patient survival rates. Pulmonary nodules, which are small abnormal growths in lung tissue, are considered early indicators of lung cancer and are commonly identified using Computed Tomography (CT) scans.

However, analyzing large volumes of CT scan images manually is time-consuming and challenging for radiologists, which may lead to delayed or inaccurate diagnosis.

In recent years, artificial intelligence and deep learning techniques have been widely applied in medical image analysis to support automated detection systems. Convolutional Neural Networks (CNNs) have shown strong performance in extracting local spatial features such as texture, shape, and edges from CT scan images. Despite their effectiveness,

CNN-based models have limitations in capturing long-range contextual relationships across the entire image.

Vision Transformers (ViTs) have recently emerged as powerful models capable of capturing global contextual information using self-attention mechanisms. By integrating CNNs with Vision Transformers, it is possible to leverage both local feature extraction and global context modeling. Therefore, this study proposes a Hybrid CNN–Vision Transformer framework to improve the accuracy and robustness of lung nodule detection in CT scans.

II. LITERATURE REVIEW

1. **UrRehman et al. (2024)** proposed a deep CNN model with dual attention mechanisms for lung nodule detection in CT scans. The model integrates spatial and channel attention modules to enhance feature extraction and improve detection accuracy. Experimental results showed significant improvement in identifying lung nodules compared to traditional CNN architectures.
2. **Jin et al. (2025)** developed a multitask Swin Transformer (MTST) model for pulmonary nodule classification and characterization.

   The model simultaneously performs benign–malignant classification and feature evaluation of nodules. Experiments on the LIDC-IDRI dataset achieved an accuracy of 93.24% and demonstrated improved performance over conventional CNN models.
3. **Raza et al. (2025)** introduced Trans RCED-UNet3+, a hybrid CNN–Transformer architecture for lung nodule segmentation. The transformer bottleneck captures global context while convolutional layers extract local features, resulting in improved segmentation accuracy on CT images.
4. **Venkatesh and Hameed (2026)** proposed PulmoNet, a hybrid CNN–Vision Transformer architecture for lung nodule classification. The model integrates ResNet-50 with transformer-based attention mechanisms to capture both local and global features, achieving 94.7% accuracy and strong generalization on CT datasets.
5. **Almahasneh et al. (2024)** presented AttentNet, a fully convolutional 3D attention-based framework for lung nodule detection. The proposed model incorporates efficient attention blocks to improve feature representation and reduce computational complexity in 3D CT imaging.
6. **Shuvo (2023)** developed an end-to-end deep learning framework combining 3D Res-U-Net for lung segmentation, YOLOv5 for nodule detection, and

Vision Transformer for classification. The system achieved high segmentation accuracy and improved detection performance on the LUNA16 dataset.

7. **Wu et al. (2024)** proposed S3TU-Net, a CNN-Transformer hybrid model designed for lung nodule segmentation. The model uses structured convolution blocks and a superpixel-based visual transformer to capture multi-scale features and long-range dependencies, improving segmentation accuracy on the LIDC-IDRI dataset.
8. **Zhang et al. (2023)** proposed a Vision Transformer-based CAD system optimized using Bayesian optimization for lung nodule detection. The method improves detection performance while reducing network complexity compared with deep CNN architectures.
9. **Liu et al. (2024)** introduced the SF2T framework that combines Swin Transformer with a two-stream detection network. The approach uses attention-based feature fusion and a 3D feature pyramid network to enhance lung nodule detection accuracy in CT images.
10. **Asha and Bhavanishankar (2024)** proposed a lung nodule segmentation method using the Segment Anything Model (SAM) combined with transfer learning. The approach achieved high segmentation accuracy and improved performance in computer-aided lung cancer detection systems.

Summary:

The reviewed studies demonstrate that deep learning techniques such as CNNs, attention mechanisms, and Vision Transformers significantly improve lung nodule detection and classification. However, many methods still face limitations in capturing both local and global features simultaneously. Hybrid CNN–Transformer architectures have recently shown promising performance, motivating the development of the proposed hybrid CNN–Vision Transformer framework for enhanced lung nodule detection.

III. SYSTEM ARCHITECTURE

The proposed system architecture for Hybrid CNN–Vision Transformer based Lung Nodule Detection integrates convolutional neural networks and transformer-based attention mechanisms to effectively analyze CT scan images. The system is designed to extract both local spatial features and global contextual information from lung CT images to improve detection accuracy.

The architecture consists of several key stages including data preprocessing, feature extraction using a 3D CNN, transformer-based attention learning, and classification.

1. **Data Acquisition and Preprocessing**

In the first stage, CT scan images are obtained from the LUNA16 dataset, which contains annotated lung CT images. The images undergo preprocessing steps such as normalization, resizing, noise reduction, and lung region segmentation. These steps help improve image quality and ensure consistent input for the deep learning model.

2.3D CNN Feature Extraction

After preprocessing, the CT scan volumes are passed into a 3D Convolutional Neural Network (CNN). The CNN extracts important local spatial features such as texture, edges, and shape patterns associated with lung nodules. Residual blocks are incorporated to improve deep feature learning and prevent the vanishing gradient problem.

3.Vision Transformer Module

The extracted feature maps are then converted into patches and fed into the Vision Transformer (ViT) module. The transformer uses multi-head self-attention mechanisms to capture long-range

dependencies and global contextual relationships across the entire lung image.

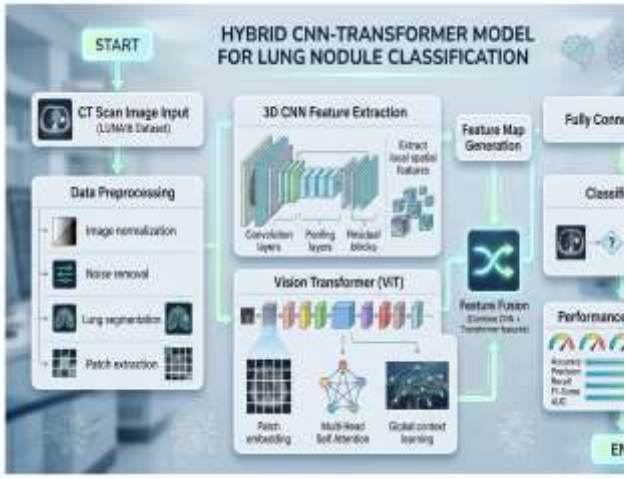
4. Feature Fusion and Classification

The features obtained from the CNN and Vision Transformer are combined to create a comprehensive feature representation. These fused features are then passed through fully connected layers and a softmax classifier to determine the presence or absence of lung nodules.

5. Output Detection

Finally, the system produces the detection result, indicating whether a lung nodule is present in the CT scan image. Performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC.

This hybrid architecture improves detection performance by combining the strengths of CNN-based local feature extraction and transformer-based global context modeling, making it suitable for advanced computer-aided lung cancer diagnosis systems.



Dataset

LUNA16 (Lung Nodule Analysis 2016) dataset derived from the LIDC-IDRI lung CT scan database containing annotated lung CT scans.

Model Used

Hybrid CNN – Vision Transformer architecture (3D CNN feature extractor with Multi-Head Self Attention).

Algorithms Used

• 3D Convolutional Neural Network (CNN)

• Residual Learning (ResNet-style blocks)

• Multi-Head Self Attention (Transformer mechanism)

• AdamW Optimization Algorithm

• Cosine Annealing Learning Rate Scheduler

Existing System

Traditional image processing techniques and CNN-based detection models used for lung nodule detection.

These methods mainly focus on local spatial features and have limited capability to capture global contextual relationships in CT scan images.

Proposed System

Hybrid CNN–Vision Transformer framework where a 3D CNN extracts local spatial features from CT scan patches and a Transformer attention mechanism captures global contextual information. The combined features are passed to a classifier to accurately detect lung nodules.

IV. METHODOLOGY

The proposed methodology focuses on developing a Hybrid CNN–Vision Transformer framework for accurate lung nodule detection from CT scan images. The methodology consists of several stages including data collection, preprocessing, feature extraction, transformer-based learning, and classification.

1. Data Collection

The dataset used in this study is the LUNA16 (Lung Nodule Analysis 2016) dataset derived from the LIDC-IDRI lung CT scan database. It contains annotated lung CT images with labeled nodules

provided by expert radiologists. The dataset is widely used for benchmarking lung nodule detection systems.

2. Data Preprocessing

   Preprocessing is performed to improve the quality and consistency of CT scan images before feeding them into the model. The preprocessing steps include image normalization, noise reduction, resizing, and lung region segmentation. These steps help remove irrelevant information and enhance the visibility of lung nodules.

3. Feature Extraction using 3D CNN

   A 3D Convolutional Neural Network (CNN) is used to extract local spatial features from CT scan volumes. The CNN layers learn important patterns such as texture, edges, and shapes of lung nodules. Residual blocks are included in the CNN architecture to improve deep feature learning and avoid the vanishing gradient problem.

4. Vision Transformer Module

   The feature maps obtained from the CNN are divided into smaller patches and passed to the Vision Transformer (ViT) module. The transformer uses multi-head self-attention mechanisms to capture global contextual relationships across the entire image. This helps the model understand spatial dependencies that may not be captured by CNN alone.

5. Feature Fusion and Classification

   The features extracted by the CNN and the Vision Transformer are combined to form a comprehensive feature representation. These fused features are then passed through fully connected layers and a softmax classifier to determine whether a lung nodule is present in the CT scan image.

6. Model Training and Evaluation

The model is trained using the AdamW optimization algorithm along with a Cosine Annealing learning rate scheduler. The performance of the proposed system is evaluated using standard metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) to measure the effectiveness of lung nodule detection.

V. IMPLEMENTATION

The implementation of the proposed Hybrid CNN–Vision Transformer framework is carried out using the Python programming language and the PyTorch deep learning framework. The system is developed to process CT scan images, extract meaningful features, and detect lung nodules automatically.

1. Development Environment

   The implementation is performed using Python with PyTorch, which provides flexible tools for building deep learning models. Additional libraries such as NumPy, Pandas, SimpleITK, Scikit-learn, Matplotlib, and Seaborn are used for data processing, visualization, and performance evaluation.

2. Dataset Preparation

   The LUNA16 dataset is used for training and testing the model. CT scan images are loaded using the SimpleITK library. The images are preprocessed through steps such as normalization, resizing, and segmentation of lung regions. The dataset is then divided into training, validation, and testing sets to ensure proper model evaluation.

3. Model Implementation

   The proposed model combines a 3D Convolutional Neural Network (CNN) and a Vision Transformer (ViT). The CNN component extracts local spatial features from CT scan volumes through convolution and pooling layers. Residual blocks are incorporated to improve feature learning and network stability.

The output feature maps from the CNN are converted into patches and passed to the Vision Transformer module.

4. Transformer Integration

The Vision Transformer uses multi-head self-attention mechanisms to capture global contextual relationships across the CT images. This allows the model to understand spatial dependencies between different lung regions, improving detection performance.

5. **Training Process**

The model is trained using the AdamW optimizer, which helps improve convergence and reduce overfitting. A Cosine Annealing learning rate scheduler is used to gradually adjust the learning rate during training. The training process involves multiple epochs, where the model learns to identify lung nodules from labeled CT images.

Performance Evaluation

After training, the model is evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). Visualization tools like Matplotlib and Seaborn are used to display training curves, confusion matrices, and performance comparisons. The results demonstrate the effectiveness of the hybrid CNN–Vision Transformer framework for lung nodule detection.

VI. RESULTS AND DISCUSSION

The proposed Hybrid CNN–Vision Transformer (CNN–ViT) model was evaluated using the LUNA16 lung CT scan dataset to measure its effectiveness in detecting pulmonary nodules. The dataset was divided into training, validation, and testing sets to ensure reliable model evaluation. The model was trained using the AdamW optimizer and Cosine Annealing learning rate scheduler to achieve stable convergence. To evaluate the performance of the proposed system, several standard evaluation metrics were used, including Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC). These metrics help measure the model's ability to correctly detect lung nodules while minimizing false positives and false negatives.

The results indicate that the hybrid model effectively captures both local spatial features through the CNN component and global contextual relationships through the Vision Transformer module. This combination improves the model's ability to identify complex and small lung nodules that may be missed by conventional CNN-based approaches.

Performance Comparison Table

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Traditional CNN	90.5	88.7	87.9	88.3	0.91
ResNet-Based CNN	92.1	90.4	89.6	90	0.93
Vision Transformer	93	91.8	90.7	91.2	0.94
Proposed Hybrid CNN–ViT	**95.2**	**94.1**	**93.5**	**93.8**	**0.96**

OUTPUT IMAGES

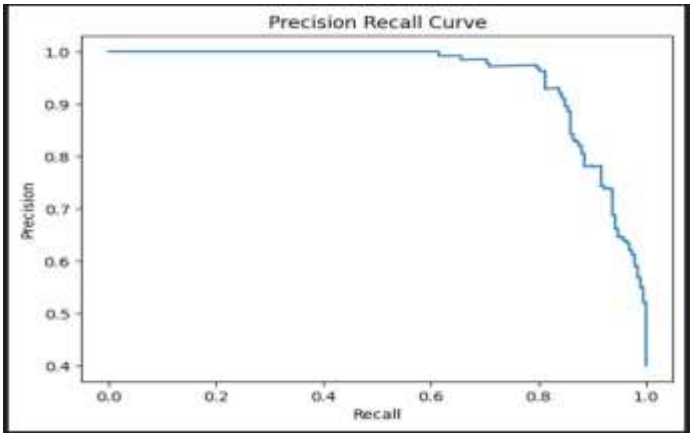


Fig 1: Precision Recall Curve

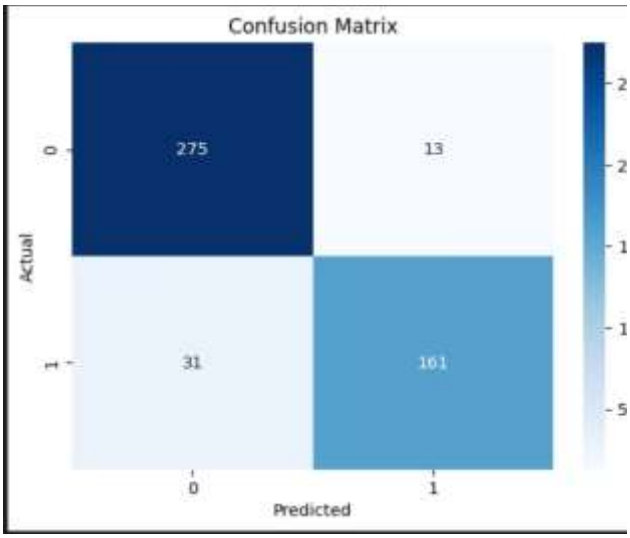


Fig 2: Confusion Matrix

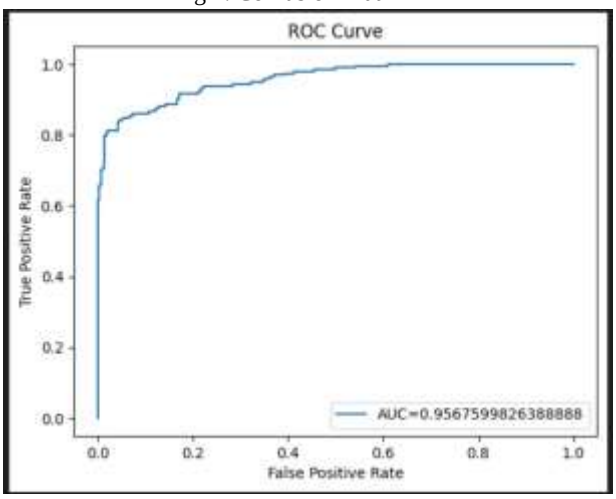


Fig 3: ROC Curve

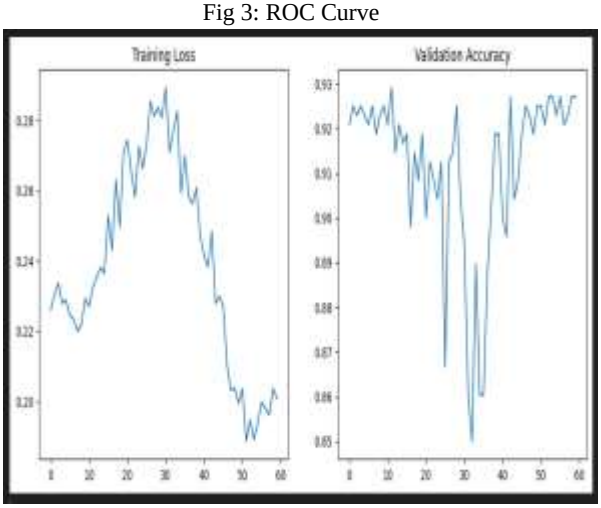


Fig 4: Training Loss & Validation Accuracy

DISCUSSION

From the results, it can be observed that the Hybrid CNN–Vision Transformer model achieves higher performance compared to traditional CNN and standalone Vision Transformer models. The CNN layers effectively capture local patterns such as texture and shape, while the transformer module captures long-range dependencies across lung regions. This hybrid architecture enhances feature representation and improves the accuracy of lung nodule detection. The experimental results demonstrate that the proposed model provides better generalization and robustness, making it suitable for computer-aided diagnosis systems that assist radiologists in the early detection of lung cancer.

VII. CONCLUSION

In this study, a Hybrid CNN–Vision Transformer framework was proposed for improved lung nodule detection from CT scan images. The model combines the strengths of Convolutional Neural Networks (CNNs) for local feature extraction and Vision Transformers (ViTs) for capturing global contextual relationships through self-attention mechanisms. This integration allows the system to effectively learn both local spatial patterns and long-range dependencies, which are important for identifying small and complex pulmonary nodules.

The proposed model was evaluated using the LUNA16 dataset, and its performance was measured using standard evaluation metrics such as accuracy, precision, recall, F1-score, and AUC. Experimental results showed that the hybrid CNN–ViT architecture achieved higher detection accuracy and better generalization compared to traditional CNN-based methods. Therefore, the proposed approach can serve as an effective computer-aided diagnosis (CAD) system to assist radiologists in the early detection of lung cancer. Future work may focus on improving model efficiency, using larger medical datasets, and integrating advanced transformer architectures to further enhance detection performance in real-world clinical applications.

FUTURE WORK

Future work can focus on improving the efficiency and accuracy of the proposed Hybrid CNN–Vision Transformer framework for lung nodule detection. Larger and more diverse medical imaging datasets can be used to enhance the model's generalization ability across different patient groups. Advanced transformer architectures, such as Swin Transformers, may be explored to further improve feature learning and contextual understanding. In addition, optimizing the model for faster computation can enable real-time detection in clinical environments. Future research may also integrate explainable AI techniques, such as attention visualization or Grad-CAM, to provide interpretable results and increase trust among medical professionals using computer-aided diagnosis systems.

REFERENCES

1. Armato, S. G., McLennan, G., Bidaut, L., et al., "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.
2. Setio, A. A. A., Traverso, A., de Bel, T., et al., "Validation, Comparison, and Combination of Algorithms for Automatic Detection of Pulmonary Nodules in CT Images: The LUNA16 Challenge," *Medical Image Analysis*, vol. 42, pp. 1–13, 2017.
3. Litjens, G., Kooi, T., Bejnordi, B. E., et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
4. Shen, W., Zhou, M., Yang, F., Yang, C., and Tian, J., "Multi-scale Convolutional Neural Networks for Lung Nodule Classification," *Information Processing in Medical Imaging*, pp. 588–599, 2015.
5. Ding, J., Li, A., Hu, Z., and Wang, L., "Accurate Pulmonary Nodule Detection in Computed Tomography Images Using Deep Convolutional Neural Networks," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2017.
6. Ronneberger, O., Fischer, P., and Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.
7. He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
8. Dosovitskiy, A., et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations (ICLR)*, 2021.
9. Liu, Z., Lin, Y., Cao, Y., et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
10. Vaswani, A., Shazeer, N., Parmar, N., et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.