A Comparative Study of CNN, Vision Transformer, and Hybrid CNN–Transformer Models for Histopathology-Based Cancer Detection

K.Anandhi[1], Dr.K.Karuppasamy[2]

[1]*PG Student, Department of Computer Science, RVS College of Engineering and Technology, Kannampalayam, Sulur, Coimbatore, TamilNadu, India;*

[2]*Professor, Department of Computer Science, RVS College of Engineering and Technology, Kannampalayam, Sulur, Coimbatore, TamilNadu, India;*

Abstract

Accurate and automated analysis of histopathology images plays a critical role in early cancer detection and clinical decision support. Recent advances in deep learning have demonstrated strong potential for improving diagnostic accuracy; however, the relative effectiveness of convolutional, transformer-based, and hybrid architectures remains an active area of research. In this study, we present a comprehensive comparative evaluation of three deep learning models such as ResNet-18, Vision Transformer (ViT), and a hybrid CNN–Transformer architecture for cancer classification using histopathological image data. All models were implemented under a unified experimental setup and evaluated using standard performance metrics, including accuracy, precision, recall, F1-score, and confusion matrix analysis. ResNet-18 serves as a robust convolutional baseline, while ViT captures global contextual relationships through self-attention mechanisms. The proposed hybrid model integrates convolutional feature extraction with transformer-based global reasoning to leverage the strengths of both paradigms. Experimental results demonstrate that the hybrid architecture consistently outperforms the individual models, achieving superior classification performance and improved generalization. These findings indicate that hybrid CNN–Transformer frameworks offer a promising direction for reliable and interpretable computer-aided cancer diagnosis from histopathology images.

Keywords: Cancer detection, Convolutional neural networks, Deep learning, Histopathology images, Hybrid CNN–Transformer model, Image classification, ResNet-18, Vision Transformer

1. Introduction

Cancer remains a leading cause of mortality worldwide, where early and accurate diagnosis plays a crucial role in improving patient outcomes[1]. Histopathological analysis of tissue biopsies is considered the gold standard for cancer diagnosis, as it enables detailed evaluation of cellular morphology and tissue architecture[2]. However, manual interpretation of histopathology slides is time-consuming and susceptible to inter- and intra-observer variability, particularly under increasing diagnostic workloads[3]. These limitations have driven the adoption of automated computer-aided diagnosis (CAD) systems to support reliable and efficient clinical decision-making.

Recent advances in deep learning have significantly improved automated histopathology image analysis[4]. Convolutional Neural Networks (CNNs) are widely used due to their ability to learn hierarchical spatial features directly from image data. Residual Networks (ResNets), in particular, introduced skip connections that alleviate vanishing gradient issues and enable stable training of deep architectures[5]. Among them, ResNet-18 provides an effective balance between classification performance and computational efficiency, making it well suited for medical imaging tasks with moderate dataset sizes.

More recently, Vision Transformers (ViTs) have been introduced to computer vision, modeling images as sequences of patches and leveraging self-attention to capture long-range dependencies and global contextual information[8]. This capability is especially valuable in histopathology, where diagnostically relevant patterns may span large tissue regions. However, ViTs generally lack strong inductive biases such as locality and translation invariance, which can limit performance when training data is limited[10].

To overcome the individual limitations of CNNs and ViTs, hybrid CNN–Transformer architectures have emerged, combining convolutional feature extraction with transformer-based global reasoning[11]. Such models aim to integrate local texture information with global contextual representations, and recent studies have reported improved robustness and generalization in medical image analysis tasks[12][13].

Motivated by these developments, this study presents a systematic comparative evaluation of three deep learning architectures—ResNet-18, Vision Transformer, and a hybrid CNN–Transformer model—for histopathology-based cancer detection. All models are trained and evaluated under a unified experimental framework using standard performance metrics, enabling a fair analysis of their relative strengths and limitations, with particular emphasis on the potential advantages of hybrid modeling for reliable cancer diagnosis.

2. Related Work

2.1 CNN-Based Histopathology Image Analysis

Convolutional Neural Networks (CNNs) have been widely adopted for histopathology image analysis due to their capability to automatically learn hierarchical feature representations from raw pixel data. Early studies demonstrated that CNN-based approaches significantly outperform traditional handcrafted feature methods for cancer detection and tissue classification tasks [1], [2]. Architectures such as AlexNet, VGGNet, and Inception have been successfully applied to classify histopathological images across multiple cancer types [3].

Residual Networks (ResNets) introduced identity-based skip connections, enabling the training of deeper networks and addressing vanishing gradient issues [4]. ResNet-based models have been extensively employed in histopathology for tumor classification, grading, and subtype prediction [5], [6]. Among these, ResNet-18 has gained particular attention due to its relatively low computational complexity and competitive performance, making it suitable for medical imaging applications with limited datasets and resource constraints [7].

2.2 Vision Transformers in Medical Imaging

Transformers were initially proposed for natural language processing and later adapted to vision tasks through the Vision Transformer (ViT) framework [8]. ViT models divide images into fixed-size patches and apply self-attention mechanisms to model global contextual relationships. This capability allows ViTs to capture long-range dependencies that are difficult for conventional CNNs to model effectively. Recent studies have explored the application of ViTs in medical imaging, including radiology, pathology, and ophthalmology [9], [10]. In histopathology image analysis, ViTs have demonstrated competitive performance, particularly when trained on large datasets or initialized using pretraining strategies [11]. However, the absence of strong inductive biases such as locality and translation invariance makes ViTs more data-intensive and computationally demanding, which can limit their effectiveness in scenarios with limited labeled medical data [12].

2.3 Hybrid CNN–Transformer Architectures

To overcome the individual limitations of CNNs and transformers, hybrid CNN–Transformer architectures have been proposed, combining convolutional feature extraction with attention-based global modeling [13]. In such models, CNN backbones are typically employed to capture local texture and morphological features, while transformer modules are used to model long-range dependencies and contextual relationships.

Hybrid architectures have shown promising results in various medical image analysis tasks, including histopathology classification and segmentation [14], [15]. By integrating local and global feature representations, these models have been reported to achieve improved robustness and generalization compared to standalone CNN or transformer-based approaches. In the context of histopathology, hybrid CNN–Transformer models are particularly effective in capturing both cellular-level details and tissue-level structural patterns, which are critical for accurate cancer diagnosis [16].

2.4 Research Gap and Motivation

Although CNNs, ViTs, and hybrid architectures have been individually explored for histopathology image analysis, direct and systematic comparisons among lightweight CNN models such as ResNet-18, pure Vision Transformer models, and hybrid CNN–Transformer frameworks under a unified experimental setting remain limited. Many existing studies employ different datasets, preprocessing strategies, and evaluation protocols, making fair comparison challenging.

Motivated by this gap, the present study implements and evaluates ResNet-18, Vision Transformer, and a hybrid CNN–Transformer model within a consistent experimental framework. By using standardized performance metrics, this work aims to provide clearer insights into the relative strengths and trade-offs of convolutional, transformer-based, and hybrid approaches for histopathology-based cancer detection.

3. Methodology

This section describes the dataset, preprocessing pipeline, model architectures, and training strategy adopted for the comparative evaluation of ResNet-18, Vision Transformer (ViT), and the proposed hybrid CNN–Transformer model for histopathology-based cancer detection.

3.1 Dataset Description

The experiments were conducted using a publicly available histopathology image dataset obtained from Kaggle [14], consisting of labeled image patches representing cancerous and non-cancerous tissue regions. The images were acquired at high magnification and exhibit substantial variability in tissue morphology, staining intensity, and cellular structure. To ensure an unbiased evaluation, the dataset was divided into training, validation, and testing subsets using a fixed split ratio while maintaining class balance across all splits. Prior to model training, all images were resized to a uniform resolution of 224×224×3 to ensure compatibility with both CNN- and transformer-based architectures, and pixel values were normalized using standard mean and standard deviation to reduce illumination and staining variability. Furthermore, data augmentation techniques including random horizontal and vertical flipping, rotation, and mild color perturbations were applied during training to enhance generalization and mitigate overfitting, which is particularly important in histopathology image analysis where tissue orientation and visual appearance can vary significantly [15].

3.2 Model Architectures

3.2.1 ResNet-18 Architecture

Residual Networks (ResNets), introduced by He et al. [5], mitigate the vanishing gradient problem in deep convolutional networks through identity-based skip connections. ResNet-18 is a lightweight yet effective variant with 18 layers, making it well suited for medical imaging applications with limited training data. The network accepts normalized histopathology patches of size 224×224×3. An initial 7×7 convolution with 64 filters and stride 2 extracts low-level cellular features, followed by a 3×3 max-pooling layer for spatial downsampling. Four residual stages (Conv2_x to Conv5_x) progressively learn hierarchical representations, ranging from basic edges and textures to complex tissue-level patterns. Skip connections ensure stable gradient flow and effective feature reuse. Global Average Pooling compresses the final feature maps into a 512-dimensional vector, which is passed to a fully connected layer and Softmax classifier to predict cancerous and non-cancerous tissue classes.

3.2.2 Vision Transformer (ViT)

The Vision Transformer (ViT) architecture reformulates image classification as a sequence modeling problem. The input image is divided into fixed-size patches, each of which is flattened and linearly projected into a D-dimensional embedding space. A learnable classification (CLS) token is prepended to the patch sequence, and positional encodings are added to retain spatial information. The resulting token sequence is processed by a stack of Transformer encoder layers comprising layer normalization, multi-head self-attention, and feed-forward networks with GELU activation, along with residual connections for training stability. After L encoder layers, the CLS

token encodes a global representation of the image and is used for final classification of cancerous versus non-cancerous histopathology regions. [9].

3.2.3 Hybrid CNN–Transformer Model

The proposed hybrid CNN–Transformer model (Fig.1) combines the complementary strengths of ResNet-18 and ViT to enhance histopathology image classification. Following preprocessing where images are resized to 224×224, normalized, and augmented where the model employs two parallel feature extraction branches. The CNN branch (ResNet-18) focuses on capturing local spatial details such as nuclei morphology and texture patterns, while the ViT branch models global contextual relationships and long-range dependencies through self-attention. The extracted 512-dimensional CNN features and 768-dimensional ViT features are concatenated to form a unified 1280-dimensional representation. This fused feature vector is refined through fully connected layers with batch normalization, ReLU activation, and dropout for regularization. Finally, a Softmax layer outputs class probabilities for cancerous and non-cancerous tissue, enabling robust and discriminative prediction.

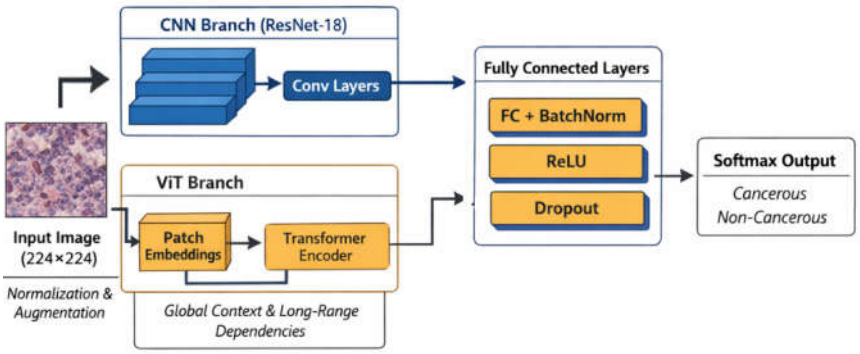

Figure 1: Proposed Model

4. Experimental Results

This section reports the quantitative performance of the evaluated models such as ResNet-18, Vision Transformer (ViT), and the proposed Hybrid CNN–Transformer on the histopathology image classification task. Performance was measured using accuracy, precision, recall, F1-score, confusion matrices, and ROC–AUC.

4.1 ResNet-18 Results:

The ResNet-18 model converged steadily across training epochs, as reflected by decreasing training and validation losses. On the test set, it achieved an accuracy of **0.8386**, precision of **0.8786**, recall of **0.7848**, and an F1-score of **0.8283**.(Fig.2). The confusion matrix indicates a noticeable number of false negatives, and the ROC curve yields an AUC of approximately **0.87**, demonstrating moderate discriminative performance.

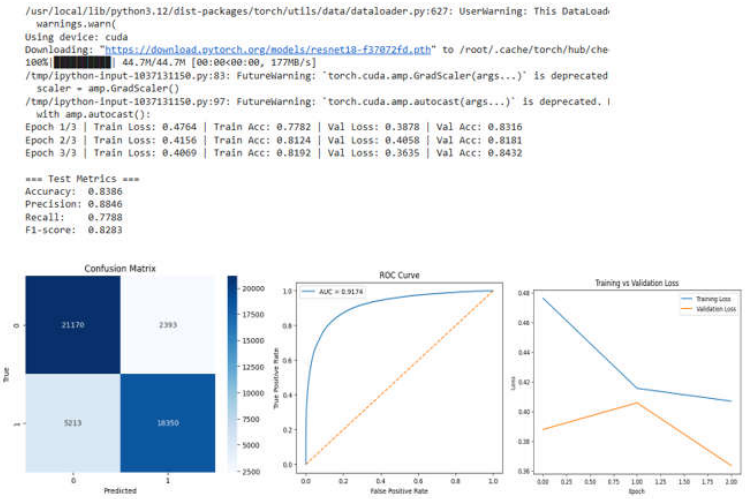

Figure 2: Resnet18 Results

4.2 Vision Transformer (ViT) Results:

The ViT model exhibited faster convergence and improved validation performance compared to ResNet-18. Test results show an accuracy of **0.9295**, precision of **0.9287**, recall of **0.9304**, and an F1-score of **0.9295**. The confusion matrix reflects balanced class predictions, and the ROC curve reports a higher AUC of approximately **0.98**, indicating strong class separability(Fig.3).

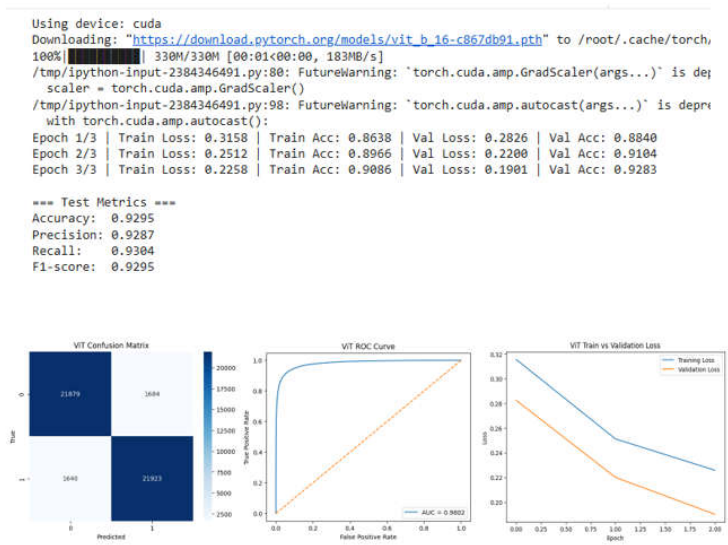

Figure 3: ViT Results

4.3 Hybrid CNN–Transformer Results:

The proposed hybrid architecture achieved the highest performance among all evaluated models.

The model demonstrated stable training behavior with consistently improving validation metrics(Fig.4a).

```
Device: cuda
Loaded ResNet checkpoint: resnet18_oscc.pth
Loaded ViT checkpoint: vit_oscc.pth
Epoch 1/3 | Train Loss: 0.2017 | Train AUC: 0.9741 | Val Loss: 0.1389 | Val AUC: 0.9893
  -> New best model saved (val AUC: 0.9893)
Epoch 2/3 | Train Loss: 0.1396 | Train AUC: 0.9870 | Val Loss: 0.1027 | Val AUC: 0.9929
  -> New best model saved (val AUC: 0.9929)
Epoch 3/3 | Train Loss: 0.1171 | Train AUC: 0.9906 | Val Loss: 0.1011 | Val AUC: 0.9937
  -> New best model saved (val AUC: 0.9937)


Evaluating on test set using best saved model (if available)...
Loaded best model: hybrid_best_by_val_auc.pth

=== Test set metrics ===
Acc: 0.9686 | Prec: 0.9834 | Rec: 0.9533 | F1: 0.9681 | AUC: 0.9950
```

Figure 4a: Hybrid Model Results

On the test set, it attained an accuracy of **0.9686**, precision of **0.9834**, recall of **0.9533**, and an F1-score of **0.9681**. The confusion matrix shows minimal misclassification, and the ROC analysis yields an AUC of **0.9950**, confirming excellent discriminative capability(Fig. 4b).

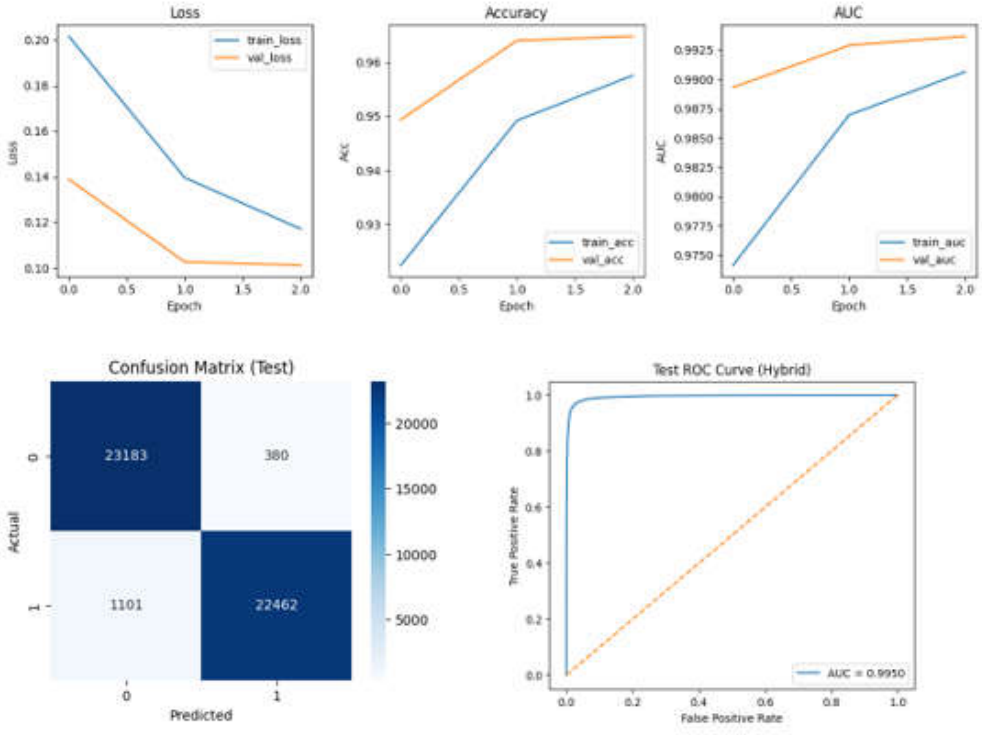

Figure 4b: Hybrid Model Results

5. Discussion

The experimental results indicate clear performance differences among the evaluated architectures. The ResNet-18 baseline, while computationally efficient, relies primarily on local feature extraction and therefore shows limitations in capturing complex tissue-level variations, leading to higher false negative rates.

The Vision Transformer improves classification performance by modeling long-range dependencies and global contextual information through self-attention, which is reflected in its substantially higher accuracy and AUC. However, transformer-based models may underutilize fine-grained spatial details that are critical in histopathological analysis.

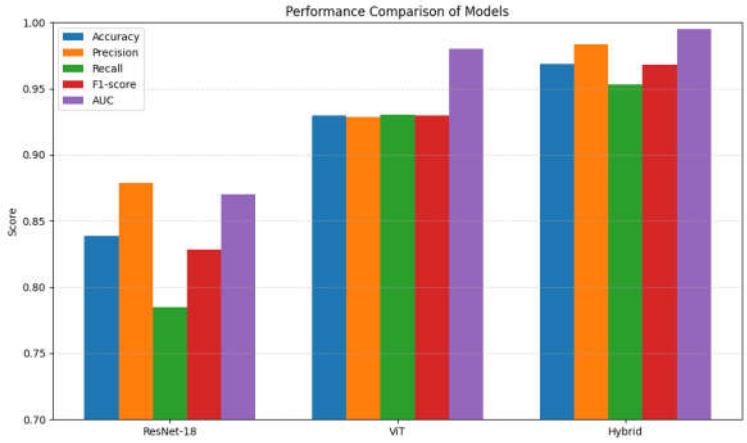

Figure 5: Performance Evaluation

The hybrid CNN–Transformer model effectively addresses these limitations by combining the complementary strengths of both architectures(Fig. 5). The CNN branch captures localized morphological patterns such as nuclei structure and texture, while the ViT branch encodes global spatial relationships across the tissue patch. The fusion of these representations results in superior classification accuracy, reduced misclassification particularly false negatives and the highest AUC. From a clinical perspective, the improved recall and F1-score of the hybrid model are particularly important, as they indicate a reduced likelihood of missed cancerous cases, underscoring the model's potential suitability for computer-aided diagnostic systems.

6. Conclusion and Future Work

This study conducted a systematic comparative evaluation of three deep learning architectures—ResNet-18, Vision Transformer, and a hybrid CNN–Transformer model—for histopathology-based cancer detection under a unified experimental framework. The results indicate that while ResNet-18 offers a computationally efficient convolutional baseline and the Vision Transformer effectively models global contextual information, the hybrid CNN–Transformer architecture achieves superior and more consistent performance across all evaluation metrics. This improvement can be attributed to the complementary integration of convolutional inductive biases for local feature learning and transformer-based self-attention for global context modeling,

enabling richer representations of both cellular morphology and tissue-level structural patterns relevant to cancer diagnosis.

Future work will extend the proposed framework toward more clinically realistic and challenging scenarios. This includes expanding the current binary classification task to multi-class histopathology analysis for distinguishing among multiple cancer subtypes. In addition, explainable artificial intelligence techniques such as Grad-CAM and transformer attention map visualization will be incorporated to enhance model interpretability and clinical transparency. Further investigations will focus on cross-dataset validation and evaluation on larger, multi-institutional datasets to assess the robustness, generalizability, and real-world applicability of the proposed models.

References

[1] World Health Organization, *Cancer Fact Sheet*, WHO, Geneva, Switzerland, 2023.
[2] J. E. Dunn, D. R. A. A. Snead, and D. J. H. Wilson, "Histopathology: The gold standard for cancer diagnosis," *Histopathology*, vol. 70, no. 1, pp. 1–3, 2017.
[3] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical Image Analysis*, vol. 33, pp. 170–175, 2016.
[4] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[6] B. Ehteshami Bejnordi et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.
[7] S. Targ, D. Almeida, and K. Lyman, "ResNet in ResNet: Generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.
[8] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.
[9] J. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2022, pp. 574–584.
[10] M. Raghu et al., "Do vision transformers see like convolutional neural networks?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
[11] Y. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
[12] X. Wang et al., "Hybrid CNN–Transformer architecture for medical image classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3811–3822, 2022.
[13] J. Huang et al., "Transformer-based feature fusion for histopathology image classification," *Pattern Recognition*, vol. 123, 2022.
[14] Kaggle, "Histopathologic Cancer Detection Dataset," 2018. Available: Kaggle dataset repository.
[15] N. Tellez et al., "Quantifying the effects of data augmentation and stain normalization in convolutional neural networks for computational pathology," *Medical Image Analysis*, vol. 58, 2019.
[17] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.