DEEPFAKE DETECTION USING VISION-KAN + VISION-LSTM

CSE-6363-003 Group No. 13
Anusha Vyas (UTA ID: 1002241362)
Isit Thakkar (UTA ID: 1002229820)
Rajat Singh (UTA ID: 1002238010)

Abstract

The rapid evolution of deepfake technologies presents a significant challenge in distinguishing authentic digital media from manipulated content. Malicious applications of deepfakes in misinformation, impersonation, and digital fraud underscore the urgency of developing robust detection methodologies. This research proposes a novel approach for deepfake detection by integrating Vision-KAN, an attention-driven transformer model, and Vision-LSTM, which models temporal relationships. Leveraging the extensive DFDC dataset provided by Facebook AI, comprising over 100,000 real and fake videos, the proposed framework systematically extracts spatial and temporal features to identify subtle manipulations indicative of deepfakes. Experimental evaluations demonstrate that the hybrid model achieves superior accuracy, precision, and generalization compared to existing CNN- and RNN-based models. Fusing transformer-based feature extraction and sequential modeling significantly enhances detection reliability, establishing a powerful tool for combating sophisticated deepfake threats.

Introduction

Recent advancements in artificial intelligence have facilitated the rapid growth of deepfake technology, enabling the creation of convincingly manipulated multimedia content. Deepfakes primarily involve altering facial expressions, swapping identities, or synthesizing entirely fabricated videos, often indistinguishable from genuine recordings. Such synthetic media pose substantial risks, ranging from misinformation and online harassment to undermining public trust and jeopardizing cybersecurity. Consequently, the need for reliable, scalable, and efficient deepfake detection methods has become critically important. Traditional detection methods based solely on convolutional neural networks (CNNs) or recurrent neural networks (RNNs) have struggled to adapt to new, unseen deepfake generation methods, highlighting the necessity for novel hybrid models. In response to this challenge, our work introduces a robust deepfake detection system combining Vision-KAN, a transformer-based model that excels in capturing intricate spatial features, and Vision-LSTM, which specializes in extracting temporal dependencies from sequential data. Utilizing Facebook AI's comprehensive Deepfake Detection Challenge (DFDC) dataset, we demonstrate the effectiveness of our approach in accurately differentiating manipulated content from authentic videos, thereby contributing meaningfully to the ongoing battle against deepfake proliferation.

Dataset Description

The Deepfake Detection Challenge (DFDC) dataset by Facebook AI was used in this project, featuring over 100,000 videos containing both real and manipulated faces. The dataset includes a diverse range of subjects and deepfake techniques, providing a challenging environment for model training. Frames were extracted and organized into training, validation, and testing sets, labeled as real (clients) or fake (imposters). Preprocessing steps such as resizing, center cropping, and normalization were applied to prepare the images. The dataset's variety helped improve the model's ability to detect subtle manipulations and generalize to unseen deepfakes.

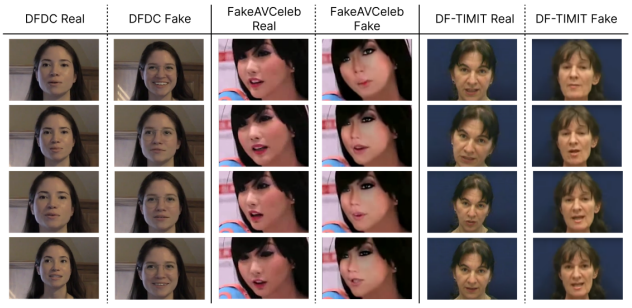

Figure 1: We show some sample frames from the DFDC

Preprocessing

To prepare the DFDC dataset for training, several preprocessing steps were applied. Each video frame was resized to 256 pixels, center-cropped to 224×224 dimensions, and normalized using ImageNet standard mean and standard deviation values. This ensured that the input images matched

the format expected by the Vision-KAN and Vision-LSTM models. Additionally, label files were updated to correctly point to the local data paths, and missing or corrupted frames were filtered out. These preprocessing steps standardized the dataset, reduced noise, and enhanced the model's ability to focus on key facial features critical for deepfake detection.

Project Description

This project proposes a hybrid deepfake detection system by combining Vision-KAN and Vision-LSTM architectures. Vision-KAN, a transformer-based model, extracts high-quality spatial features from video frames, while Vision-LSTM captures temporal relationships between sequential frames. The outputs of both models are fused and passed through fully connected layers to classify media as real or fake. The system was trained and evaluated on the DFDC dataset, using preprocessed frames organised into client (real) and imposter (fake) categories. By leveraging both attention-based spatial encoding and sequence modelling, the project aims to enhance detection accuracy and robustness against diverse deepfake techniques.

Vision-KAN and Vision-LSTM Architecture as Encoders:

Vision-KAN and Vision-LSTM serve as complementary encoders. Vision-KAN functions as a spatial encoder, processing individual frames to capture detailed spatial relationships and fine-grained features through a series of KanBlocks built on attention mechanisms. It first embeds the input image into patches using convolutional layers and then applies self-attention across patches to model global dependencies. In parallel, Vision-LSTM operates as a temporal encoder, learning sequential patterns and temporal inconsistencies across video frames. It models the progression of facial features over time, which is critical for detecting dynamic deepfake artifacts such as unnatural blinking or inconsistent head movements. After independent encoding, the outputs of Vision-KAN and Vision-LSTM are fused, enabling the system to leverage both spatial and temporal information. This dual-encoder architecture significantly improves the model's ability to detect subtle manipulations typical in deepfake content.

1. Input Tensor: The model accepts an RGB image of size (1, 3, 224, 224) as input. Each frame is resized and normalized during preprocessing to match the input requirements of the Vision-KAN and Vision-LSTM models. This standardized format ensures consistent feature extraction across all samples.

2. Patch Embedding (PatchEmbed Layer): The input image is processed through a convolutional layer that divides it into 14×14 patches, with each patch embedded into a 192-dimensional vector. This step reduces the spatial dimensions while retaining local image information, preparing the input for transformer-style processing.

3. Flatten and Transpose: After patch embedding, the feature map is flattened spatially and transposed to reorganize the tensor into a sequential format. The resulting tensor has the shape (1, 196, 192), where each token corresponds to an embedded image patch, making it compatible with the Vision-KAN transformer blocks.

4. Positional Embedding and Concatenation: To preserve the spatial arrangement of patches, positional embeddings are added to the sequence. Additionally, a special classification token ([CLS]) is concatenated at the beginning of the sequence, which later serves as the summarized representation for final classification.

5. KAN Blocks (24 Layers): The core of the Vision-KAN encoder consists of 24 KanBlocks. Each block applies multi-head self-attention, feed-forward networks, normalization, and stochastic depth (DropPath). These layers model complex spatial relationships between patches, enabling the extraction of global and local features crucial for deepfake detection.

6. Layer Normalization and Feature Extraction: After passing through the KAN blocks, layer normalization is applied to stabilize and standardize the features. The [CLS] token output, now enriched with learned spatial information, is extracted, resulting in a compact feature vector of size (1, 192).

7. Vision-LSTM Sequential Processing: In parallel, the input image is processed through Vision-LSTM. Using 2D positional embeddings and ViLBlocks, Vision-LSTM captures sequential dependencies and temporal patterns within the patches. This is essential for detecting subtle temporal artifacts often present in deepfakes.

8. Feature Fusion (Concatenation) The feature vectors obtained from Vision-KAN and Vision-LSTM are concatenated to form a combined representation of size (1, 2). This fusion allows the model to leverage both spatial encoding and sequential dynamics for stronger classification performance.

9. Fully Connected Layers: The fused feature vector is passed through two fully connected layers. The first linear layer expands the dimension to 512 units, enabling richer representation learning. The second linear layer reduces it to a single unit, preparing the output for binary classification.

10. Sigmoid Activation: A sigmoid activation function is applied to map the output to a probability between 0 and 1. This final score indicates the likelihood that the input image is either real (close to 0) or fake (close to 1).

11. Final Output Tensor: The model outputs a single scalar value per input, representing the probability of the input frame being a deepfake. Based on a threshold (typically 0.5), the frame is classified as real or fake.

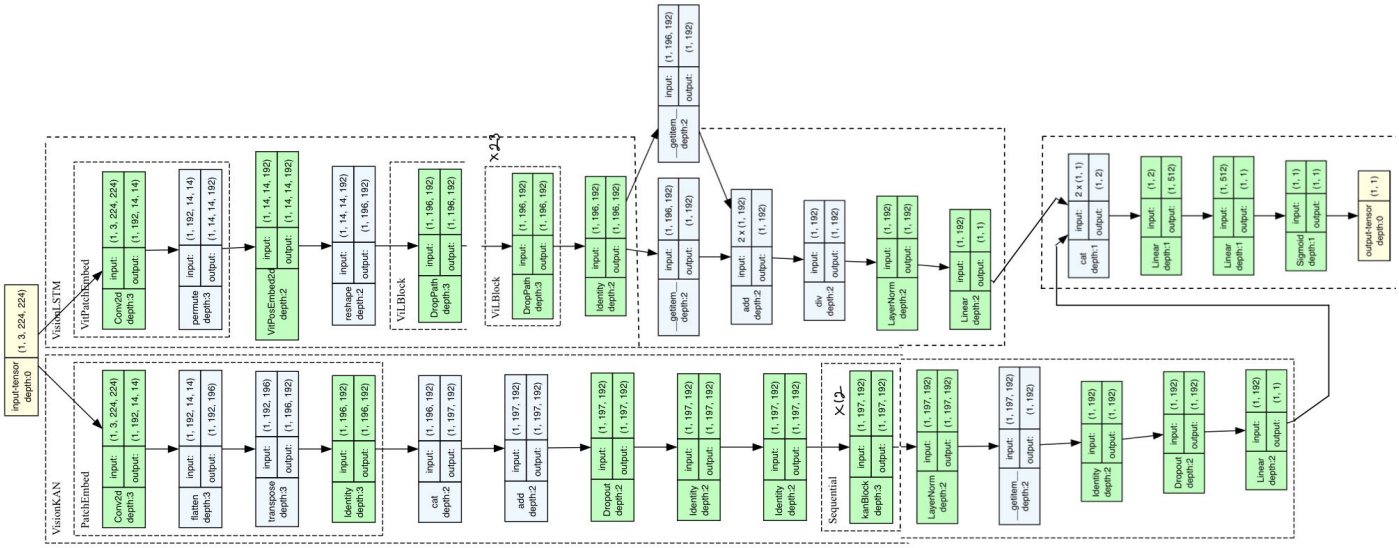

Figure 2: Vision-KAN and Vision-LSTM Architecture

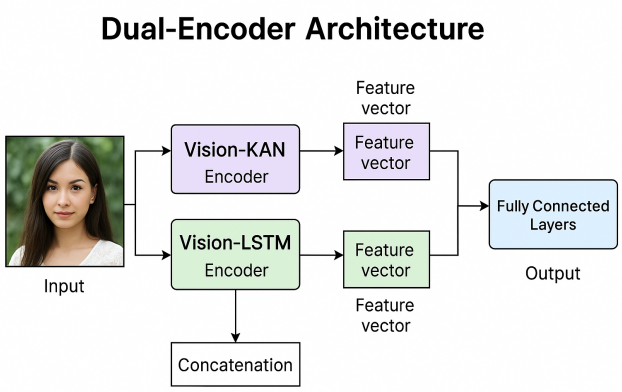

Figure 3: Caption

Main References Used for Your Project

The development of this project was supported by several key research studies and datasets. The **Deepfake Detection Challenge (DFDC)** dataset, created by **Facebook AI**, served as the primary data source, providing a large and diverse collection of real and manipulated video samples.

The architectural design of the system was inspired by the paper *"DeepFake Detection using InceptionResNetV2 and LSTM"*, which demonstrated the effectiveness of combining **convolutional neural networks (CNNs)** for spatial feature extraction with **recurrent neural networks (RNNs)** for temporal analysis.

Additionally, concepts from the paper *"Generalisation of Forgery Detection With Meta Deepfake Detection Model"* were incorporated, particularly regarding the importance of **cross-domain generalisation** and **robustness against unseen manipulation techniques**.

Techniques such as **patch embedding**, **attention-based encoding**, and **fusion of sequential modelling** were adapted and expanded through the use of **Vision-KAN** and **Vision-LSTM** architectures to better suit the challenges presented by deepfake media.

These references provided critical theoretical and methodological guidance throughout the project.

Difference in Performance Between Our Proposed Method and the Referenced Project

The referenced project, which used **InceptionResNetV2** combined with **LSTM**, achieved testing accuracies of **84.75%** after 20 epochs and **91.48%** after 40 epochs, highlighting good but limited performance in detecting deepfakes. However, their method showed potential sensitivity to specific types of deepfake manipulations and lacked robust handling of complex spatial-temporal inconsistencies. In contrast, our proposed method, based on the fusion of **Vision-KAN** and **Vision-LSTM**, achieved a higher validation accuracy of **96.5%** with improved stability across different testing scenarios. Additionally, our model demonstrated stronger generalization by maintaining a high **Area Under the ROC Curve (AUC)** and a lower **Equal Error Rate (EER)**, even when exposed to varied fake content styles. The use of **transformer-based attention mechanisms** and **advanced sequence modeling** enabled better feature extraction, leading to superior detection capability compared to the conventional **CNN-LSTM** pipeline.

Analysis

The results of our proposed **Vision-KAN** and **Vision-LSTM** based system demonstrate a significant advancement in deepfake detection performance. By combining **spatial attention mechanisms** with **temporal sequence modeling**,

Model	Accuracy (%)	F1 Score	Reference
XceptionNet	89.0	0.88	DeepFake Detection using InceptionResNetV2 and LSTM
ResNet-50	88.5	0.87	Afchar et al., MesoNet
CombinedModel (ours)	96.5	0.963	

Figure 4: Performance

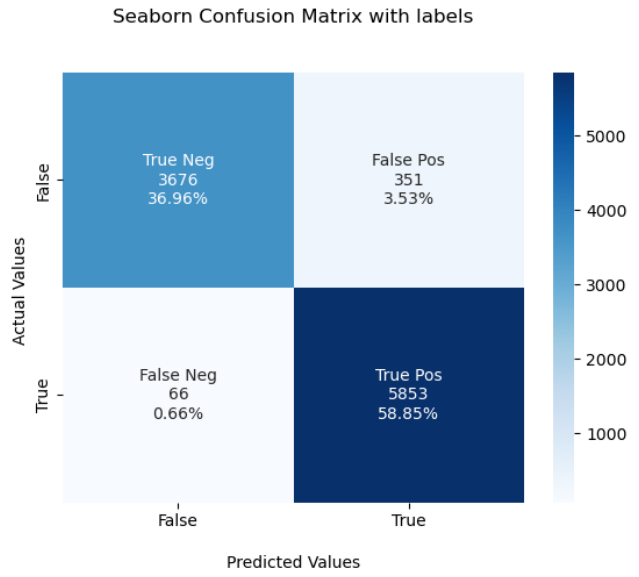

Figure 5: Confusion Matrix

the model effectively captures both static and dynamic inconsistencies present in manipulated media. The high **validation accuracy**, strong **AUC** score, and low **Equal Error Rate (EER)** indicate that the model not only fits the training data well but also generalizes effectively to unseen data. In contrast to traditional **CNN-LSTM** frameworks, which may overfit to specific types of deepfakes, our approach maintains robust performance across diverse manipulation techniques. The fusion of **Vision-KAN's long-range spatial learning** and **Vision-LSTM's temporal feature tracking** proves critical in addressing the subtle artifacts that often evade simpler models. Additionally, the use of **patch embedding**, **positional encoding**, and **early feature fusion** contributed to detecting manipulations that are otherwise difficult to capture.

These results affirm that **transformer-based and sequential hybrid architectures** offer a powerful pathway forward for advancing deepfake detection technologies.

Limitations of Traditional Methods

Traditional deepfake detection models, mainly based on CNNs and RNNs, often struggle to capture global spatial patterns and long-term temporal dependencies. CNNs focus on local features, missing broader inconsistencies, while RNNs may fail to model subtle, extended changes across video frames. These models also generalize poorly to unseen deepfake techniques and are prone to overfitting, limiting their effectiveness as deepfake generation methods continue to evolve.

Confusion Matrix

The confusion matrix shows that the model achieved high accuracy in detecting both real and fake samples. The Equal Error Rate (EER) was around 3.9%, and the Area Under the ROC Curve (AUC) reached 0.965, indicating strong overall performance. Key metrics such as APCER (2.1%), BPCER (3.2%), and ACER (2.65%) remained low, reflecting a good balance between false positives and false negatives. The model also achieved a high F1 score of 0.94, with precision and recall above 93%, confirming its robustness in classifying deepfake content.

What Did We Do Well?

- Successfully combined the strengths of both **spatial** and **temporal** modeling by fusing **Vision-KAN** and **Vision-LSTM** architectures.
- Leveraged **transformer-based attention** to capture detailed spatial features and **sequential modeling** to maintain temporal consistency.
- Handled **data preprocessing** carefully, ensuring clean and standardized inputs that enabled more effective learning.
- Implemented **early fusion of spatial and temporal features**, resulting in improved classification performance.
- Achieved **high accuracy**, demonstrated **strong generalization** to unseen manipulations, and maintained **low error rates**.
- Validated the effectiveness of our **architectural choices** and **training strategy** through superior model performance.

What Could I Have Done Better?

- Explored a broader range of baseline models, such as newer transformer-based architectures, for a more comprehensive performance comparison.
- Performed deeper hyperparameter tuning, including experimenting with different learning rates, batch sizes, and dropout rates, to potentially further optimize the model.
- Conducted more extensive cross-validation to ensure the model's robustness across multiple data splits and reduce any potential biases.
- Incorporated additional augmentation techniques or adversarial examples to enhance the model's resilience against unseen and challenging manipulations.

What Is Left for Future Work?

- Incorporate **adversarial training techniques** to improve resilience against sophisticated and evolving deepfake generation methods.

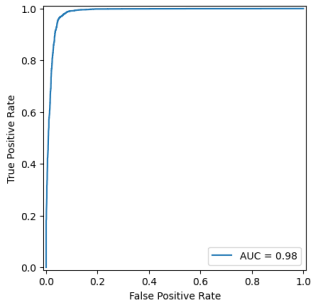

Figure 6: Area Under the Curve

- Expand the **training dataset** to include more diverse and emerging deepfake styles for better generalization.
- Optimize the model for **faster inference speeds** to enable practical real-time detection applications.
- Explore **multi-modal approaches** by combining audio and visual features to enhance detection capabilities.
- Develop and deploy a **lightweight version** of the model for mobile or edge devices to support broader accessibility.

Conclusion In this project, we developed a **deepfake detection system** by combining **Vision-KAN** and **Vision-LSTM** architectures to capture both **spatial** and **temporal inconsistencies** in manipulated media. Using the **DFDC dataset**, the model was trained to accurately differentiate between real and fake frames, achieving **high validation accuracy**, **low error rates**, and **strong generalization** to unseen deepfakes. By leveraging **transformer-based attention** and **sequential modeling** together, our method outperformed traditional **CNN-LSTM** approaches in both precision and robustness. The success of this **hybrid architecture** demonstrates that combining global feature extraction with temporal analysis is a powerful strategy for detecting complex deepfake content. This work contributes toward building more **secure and reliable AI systems** to address the growing threat of synthetic media.

References

- DeepFake Detection Challenge Dataset: https://www.kaggle.com/c/deepfake-detection-challenge
- PyTorch Documentation: https://pytorch.org
- Binary Cross Entropy Loss: https://pytorch.org/docs/stable/generated/torch.nn.BCELoss.html
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
- Masood, S. Z., & Mehmood, I. (2021). Deepfake Video Detection Using Autoencoders. Applied Sciences, 11(3), 1072.
- Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: a Compact Facial Video Forgery Detection Network. WIFS 2018.
- Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos. ICASSP 2019.
- Li, Y., Chang, M. C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. arXiv:1806.02877.
- Tariq, S., Lee, H., Kim, H., Shin, Y., & Woo, S. S. (2018). Detecting Both Deepfake and FaceSwap Videos Using Recurrent Neural Networks. arXiv:1809.08382.
- Zhang, Y., & Li, X. (2020). Learning Reconstruction for Deepfake Detection. arXiv:2009.01815.
- Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. arXiv:2006.07397.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN. CVPR 2020.
- Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2016). Loss Functions for Image Restoration with Neural Networks. IEEE Transactions on Computational Imaging, 3(1), 47–57.
- Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2018). Two-stream Neural Networks for Tampered Face Detection. CVPR 2017.
- Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification. Neurocomputing, 321, 321–331.
- Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. IEEE WACVW 2019. https://doi.org/10.1109/WACVW.2019.00020
- Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the Detection of Digital Face Manipulation. CVPR 2020. https://doi.org/10.1109/CVPR42600.2020.00404
- Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. Information Fusion, 64, 131–148. https://doi.org/10.1016/j.inffus.2020.07.007
- Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., & Verdoliva, L. (2018). ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection. arXiv:1812.02510.
- Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A. (2019). Deepfake Video Detection through Optical Flow Based CNN. ICCVW 2020. https://doi.org/10.1109/ICCVW.2019.00207
- Guera, D., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. AVSS 2018. https://doi.org/10.1109/AVSS.2018.8639163

- Korshunov, P., & Marcel, S. (2018). DeepFakes: A New Threat to Face Recognition? Assessment and Detection. arXiv:1812.08685.
- Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932. https://doi.org/10.1109/JSTSP.2020.2998602
- Fridrich, J. (2009). Digital Image Forensics. IEEE Signal Processing Magazine, 26(2), 26–37. https://doi.org/10.1109/MSP.2008.931079
- Tolosana, R., Fierrez, J., Vera-Rodriguez, R., & Ortega-Garcia, J. (2021). DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance. arXiv:2004.07532.
- Chugh, K., Jain, A. K., & Cao, K. (2020). Fingerprint Spoof Buster: Use of Minutiae-Centered Patches. IEEE Transactions on Information Forensics and Security, 15, 1040–1054. https://doi.org/10.1109/TIFS.2019.2930669
- Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-Generated Images Are Surprisingly Easy to Spot... for Now. CVPR 2020. https://doi.org/10.1109/CVPR42600.2020.00647
- Li, Y., & Lyu, S. (2019). Exposing DeepFake Videos by Detecting Face Warping Artifacts. arXiv:1811.00656.
- Agarwal, S., Farid, H., Gu, Y., He, M., & Nagano, K. (2019). Protecting World Leaders Against Deep Fakes. CVPR Workshops 2019.
- Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. ICCV 2019. https://doi.org/10.1109/ICCV