Automatic Segmentation And Labelling of Objects in Video

Main Project Presentation: First Review

Guided by:
Dr. Deepa V.

Presented by:
Batch 4

Ajay T Shaju, SJC20AD004
Emil Saj Abraham, SJC20AD028
Justin Thomas Jo, SJC20AD046
Vishnuprasad KG, SJC20AD063

Outline

- Introduction
- Problem Statement
- Application
- Literature Survey
- Block Diagram
- Data Collection
- Work Done So Far
- Work To Be Done
- Conclusion
- References

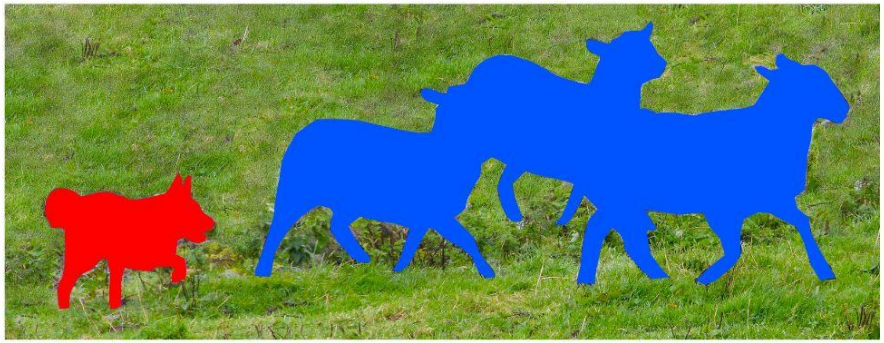
Introduction

- **How does a Self Driving Car study the environment.**
- **How does a wheeled-robot navigates through its surroundings?**
- We need to train them on **huge sets of labelled data, a huge manual work.**

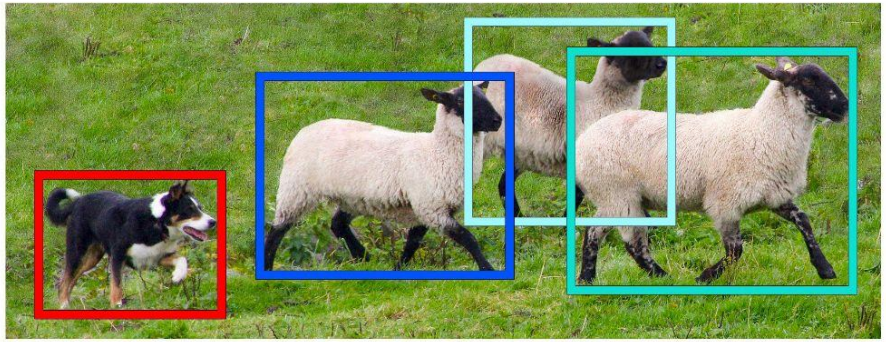
This is Done by Automating Segmentation And Labelling of Video Data.

- Object segmentation: Partitioning an image into multiple regions or segments, where each segment corresponds to a distinct object or region of interest.
- Labeling: Assigning a category or label to each segmented object.

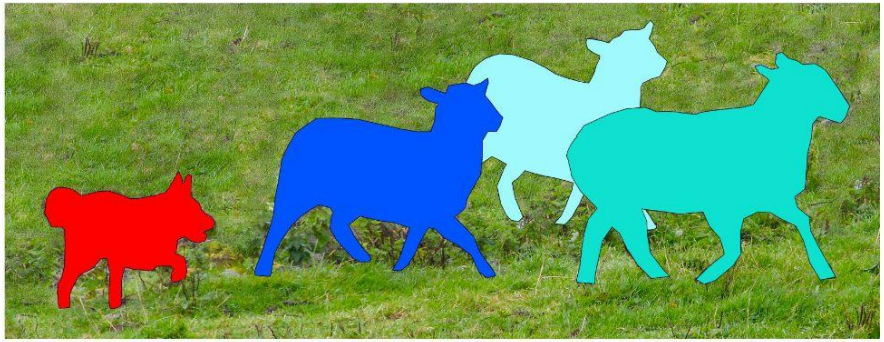
Image Recognition



Semantic Segmentation



Object Detection



Instance Segmentation

Problem Statement

- **Manual Annotation Overload** – time-consuming and resource-intensive processes.

- **Real-time Application Hurdles** – autonomous vehicles and real-time surveillance, demand swift and precise object identification in dynamic environments.

- To make data of high quality from areas of fast movements
    - **Traffic** - high speed cameras, advanced sensors, computer vision and image processing.
    - **Sports** - player tracking system, real time data analysis.

Application

- **Identifying** and **tracking objects** within a video stream in **surveillance systems**, **autonomous vehicles** and **robotics.**
- **Retrieval** of **video content** based on specific **objects, scenes, or actions.**
- Identify **regions of interest** and allocate **varying levels** of **Video compression.**
- Identify and analyze **anatomical structures** or **abnormalities** in Medicine.
- Track and analyze **movements** and **tactical strategies** in sports videos.
- **Psychoanalysis** of **human behavior** in social interactions.
- **Engaging** and **responsive** experiences in **Entertainment & Gaming.**

Literature Review

[1] Yao, Rui, et al. **"Video object segmentation and tracking: A survey."** ACM Transactions on Intelligent Systems and Technology (TIST) 11.4 (2020): 1-47.

- Solved the difficulties in handling fast motion, out-of-view, and real-time processing by **video object segmentation and tracking(VOST)**.
- Tried with different learning methods like unsupervised VOS, semi-supervised VOS, interactive VOS, and segmentation-based tracking methods.
- **Gap:** Affected by low resolution, motion blur. Segmentation is **not flexible with object shape**.
- **Future Scope**: Multi-camera video object segmentation and tracking, 3D video object segmentation and tracking.

Literature Review

[2] Yiwen Wang; Ye Lyu; Yanpeng Cao; Michael Ying Yang **“Deep Learning for Semantic Segmentation of UAV Videos”** IEEE International Symposium on Geoscience and Remote Sensing (2019)

- Proposed model combines **FCN** & **LSTM** for segmentation.
- FCN segments each frame individually.
- LSTM acts as the **post processing method** that uses temporal information of consecutive frames.
- **Gap:** Noise in the frame reduces accuracy. Resized images shows **low performance**.
- **Future Scope:** Noise reduction algorithms and better hardware equipment could significantly **increase accuracy**.

Literature Review

[3] Yang, Linjie, Yuchen Fan, and Ning Xu. "**Video instance segmentation.**" Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

- Presented the first large scale dataset, **YouTube-VIS**, for video instance segmentation
- Proposed **MaskTrack R-CNN** for video instance segmentation.
- **Gap:** unable to associate objects due to object occlusions and fast motion.
- **Future Scope:** object detection with spatial-temporal features, end-to-end trainable matching criterion, and incorporating motion information for better recognition.

Literature Review

[4] L.Zhang, Y.Lu, "**Video Object Segmentation by Latent Outcome Regression**" IEEE Access, Vol.8, pp: 30355-30367, Feb 2020.

- Proposed an **unsupervised** approach where weights and outcomes were optimized simultaneously in an **iterative** manner.
- **Weight** learning and **segmentation** inference modules **collaborate** and improve the quality.
- Weights were **adjusted** or **adapted** based on the specific characteristics of the **outcome.**
- **Gap : additional time** needed for the **aggregation** algorithm to combine the results, making this slower, lacks consistency and robustness.
- **Future Scope:** Implementation of **DL** techniques, **parallelization** strategies.

Literature Review

[5] H.Liang, L.Liu, Y.Bo, C.Zuo, "**Semi-Supervised Video Object Segmentation Based on Local and Global Consistency Learning**", IEEE Access, Vol.9, pp: 127293-127304, Sep 2021.

- Used more **unlabeled** frames to improve the **robustness** and **generalization.**
- Account the **local** and **global** information of the video.
- Reduced **complexity** and **memory consumption.**
- Great **segmentation** effect and **high prediction** speed.
- **Gap : accuracy** of the model was **insufficient** (68% approx.), Tested on **very few samples.**
- **Future Scope:** developing **fine-tuning strategies** , incorporating **more** range of **scenes**, **objects**, and **scenarios ,** Lightweight model for **real-time** processing.

Literature Review

[6] F.Jiaqing, Z.Kaihua, Z.Yaqian, L.Qingshan, "**Unsupervised Video Object Segmentation via Weak User Interaction and Temporal Modulation**", Chinese Journal of Electronics, Vol.32, No.3, May 2023.

- Uses a simple human-made rectangle annotation in the initial frame as prior information to guide the segmentation process.
- Utilizes **ETM module** and **CTM module**, to incorporate temporal information and improve the accuracy of segmentation.
- **Gap :** the inability to distinguish the contour of tiny components in high-speed sequences and the limited learning ability of the model in discriminating background areas.
- **Future Scope:** by utilizing advanced temporal modulation techniques, the proposed approach can be used in other related tasks, such as **video object tracking** and **re-identification**.

Literature Review

[7] A.Ilioudi, A.Dabiri, Ben J.Wolf, B.De Schutter, "**Deep Learning for Object Detection and Segmentation in Videos: Toward an Integration With Domain Knowledge**", IEEE Access, Vol.10, pp: 34562-34576, Mar 2022.

- Comparison of two stage approach and one stage approach in CNN-based image object detectors.
- Methods used - optical flow, tracking, LSTM, GRU, **self attention mechanisms**, generative learning.
- **Gap:** Data scarcity, generalizability, indistinguishable and lack of reasoning, **rationalized output from the techniques.**
- **Future Scope**: **Out-of-distribution generalization**, deep learning systems with causal structures, effective representation learning with few or no labeled data, **adaptation in time-varying environments**, multimodal learning

Literature Review

[8] Usman A.Usmani, J.Watada, J.Jaafar, Izzatdin A.Aziz, A.Roy, "**A Reinforcement Learning Based Adaptive ROI Generation for Video Object Segmentation**", IEEE Access, Vol.9, pp: 161959-161977, Dec 2021.

- ZVOS - Zero Shot Video Object Segmentation.
- A single, end-to-end RL framework; **Deep Deterministic Policy Gradient (DDPG)** algorithm; group co-attention mechanism
- **Gap:** Difficulty in correctly distinguishing the primary objects from the complex background when there is **no prior object** present.
- **Future Scope**: **Video saliency detection** and optical flow estimates, more powerful co-attention mechanisms, idea of meta learning, the algorithm can be tested for the detection of the primary objects in more complex scenarios.

Literature Review

[9] M.Shahid, John J.Virtusio, YH Wu, YY Chen, M.Tanveer, K.Muhammad, KL Hua, "**Spatio-Temporal Self-Attention Network for Fire Detection and Segmentation in Video Surveillance**", IEEE Access, Vol.10, pp: 1259-1275, Jan 2022.

- A novel two-stage fire-detection approach - implements spatial-temporal network.
- Uses self-attention on **Spatio-Temporal features** that are discriminative of fire, enabling our network to produce superior segmentation masks to use as region proposals.
- Constructed a video dataset containing manually generated ground-truth segmentation masks.
- **Gap:** Fires can have an **arbitrary shape, size, and even location** on the image, making it harder to learn; No large datasets are available containing fire and ground-truth segmentation masks, adding another layer of complexity.
- **Future Scope**: Making a light-weight model to run on devices with **computational or memory constraints.**

Literature Review

[10] H.Park, J.Yoo, G.Venkatesh, N.Kwak, "**Adaptive Template and Transition Map for Real-Time Video Object Segmentation**", IEEE Access, Vol.9, pp: 116914-116926, Aug 2021.

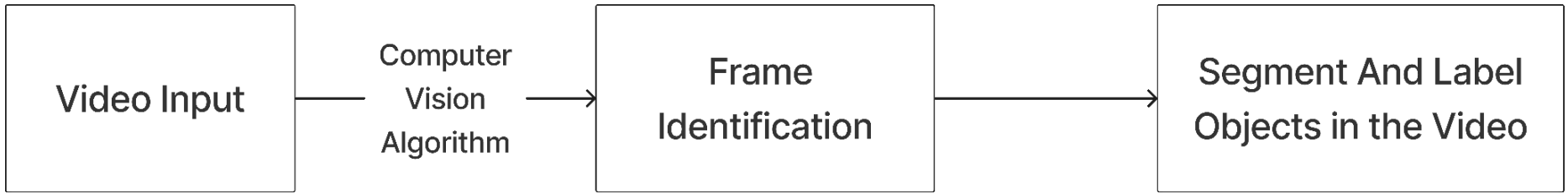
- A lightweight semi-VOS model based on two template matching methods: **short-term matching for localization and long-term matching for fine mask generation.**
- A proposed transition map for **auxiliary loss** to learn how to correct mis-estimated pixels from previous frames for current frames to prevent **error propagation**.
- **Gap:** When the target object disappears due to occlusion, many regions in the previous mask become zeros, thus we cannot give attention clues for the next frame, and the model has to find the target in a broader range, which in turn causes **performance degradation**.
- **Future Scope**: Designing a more robust method to handle problems in environments affected with **severe occlusions**.

Objectives

- Focus on **robustness**.
- Achieve real-time performance on **real-world videos**.
- Evaluate on **benchmark datasets**.
- Create **user-friendly** tool.

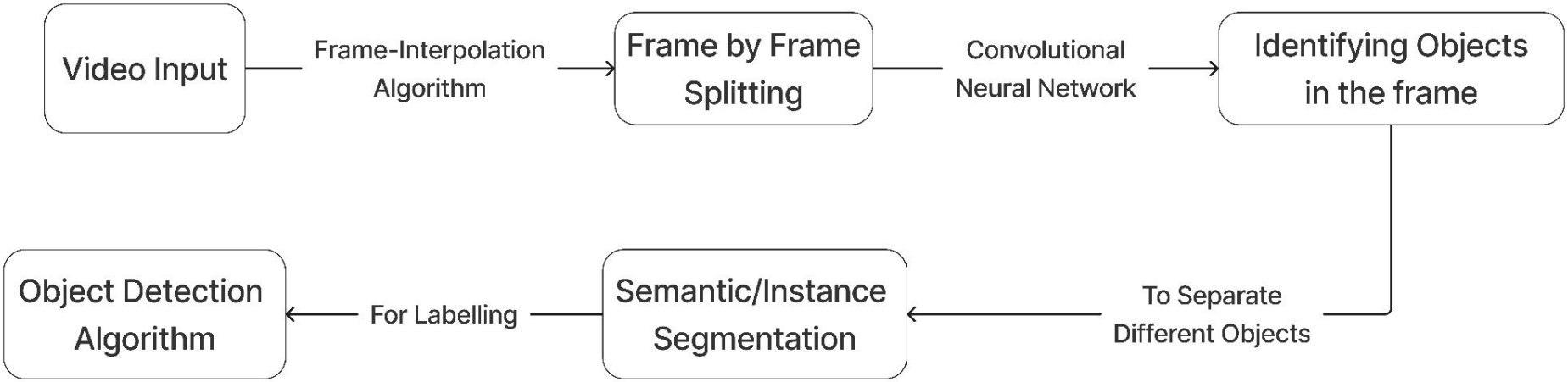
Block Diagram

High-Level Overview Of The System



Block Diagram

Detailed View Of The System

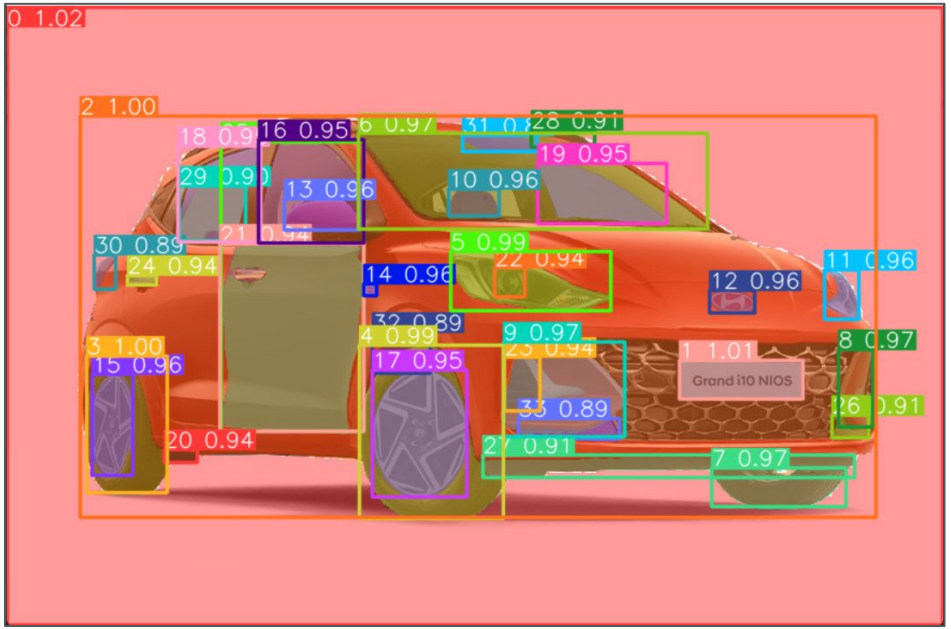


Data Collection

- Primary data - Video files of any formats.
- **Data Collection Methods:**
  - Manual data collection by taking videos ourselves.
  - From major datasets like **CamVid**(Cambridge-Driving Labelled Video Database) or **DAVIS 2016**.
    - Real-World Urban Videos – highly relevant for autonomous driving tasks.
    - Diverse Environmental Conditions – lighting and colors.
    - Benchmark dataset for video segmentation algorithms.

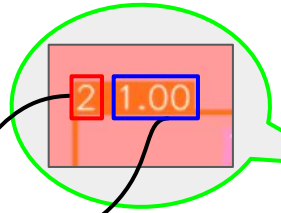
Work done so far

Source Image



Output Image after segmentation and annotation

Work done so far

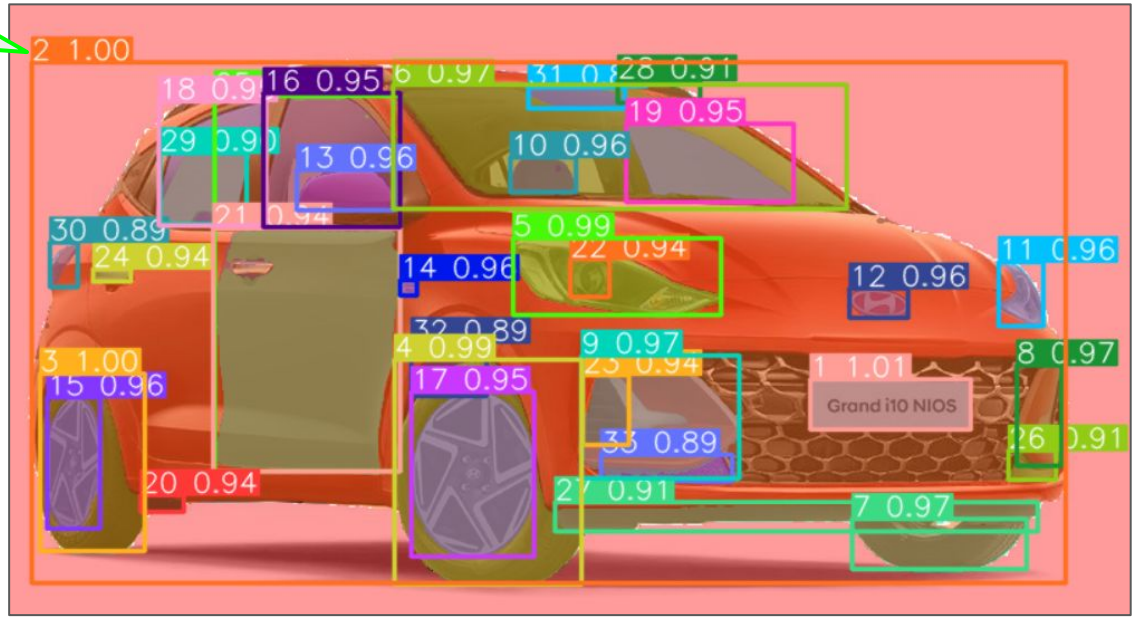


Class Label

Confidence

0	person
1	bicycle
2	car
3	motorcycle
4	airplane

Class Label of Items in COCO Dataset



Work to be done

- **Data Preprocessing**: Clean and preprocess the data, addressing issues like noise, resolution disparities, and format standardization for uniformity.
- **Model Selection and Development**:
  - **Research and Analysis**: Explore state-of-the-art machine learning models suitable for video object detection and segmentation.
  - **Model**: Implement or adapt models, potentially considering deep learning architectures such as variations of convolutional neural networks (CNNs) or recurrent neural networks (RNNs) for temporal information.
- **Real-time Implementation and Optimization**: Adapt the trained model for real-time applications, optimizing for speed and efficiency without compromising accuracy.

Conclusion

- The project **targets a major challenge** in different areas like **self-driving** vehicle development, sports etc..
- Adopting state-of-the-art algorithms to **automate segmentation and labeling**.
- Acceleration of progress towards **fully autonomous** vehicles by revolutionizing the data labeling process.

References

[1] Yao, Rui, et al., "**Video object segmentation and tracking: A survey**", ACM Transactions on Intelligent Systems and Technology (TIST) 11.4, April 2020.

[2] Yiwen Wang; Ye Lyu; Yanpeng Cao; Michael Ying Yang**, "Deep Learning for Semantic Segmentation of UAV Videos",** IEEE International Symposium on Geoscience and Remote Sensing July 2019.

[3] Yang, Linjie, Yuchen Fan, and Ning Xu, "**Video instance segmentation**" Proceedings of the IEEE/CVF International Conference on Computer Vision. May 2019.

[4] L.Zhang, Y.Lu, "**Video Object Segmentation by Latent Outcome Regression**" IEEE Access, Vol.8, pp: 30355-30367, Feb 2020.

[5] H.Liang, L.Liu, Y.Bo, C.Zuo, "**Semi-Supervised Video Object Segmentation Based on Local and Global Consistency Learning**", IEEE Access, Vol.9, pp: 127293-127304, Sep 2021.

[6] F.Jiaqing, Z.Kaihua, Z.Yaqian, L.Qingshan, "**Unsupervised Video Object Segmentation via Weak User Interaction and Temporal Modulation**", Chinese Journal of Electronics, Vol.32, No.3, May 2023.

[7] A.Ilioudi, A.Dabiri, Ben J.Wolf, B.De Schutter, "**Deep Learning for Object Detection and Segmentation in Videos: Toward an Integration With Domain Knowledge**", IEEE Access, Vol.10, pp: 34562-34576, Mar 2022.

[8] Usman A.Usmani, J.Watada, J.Jaafar, Izzatdin A.Aziz, A.Roy, "**A Reinforcement Learning Based Adaptive ROI Generation for Video Object Segmentation**", IEEE Access, Vol.9, pp: 161959-161977, Dec 2021.

[9] M.Shahid, John J.Virtusio, YH Wu, YY Chen, M.Tanveer, K.Muhammad, KL Hua, "**Spatio-Temporal Self-Attention Network for Fire Detection and Segmentation in Video Surveillance**", IEEE Access, Vol.10, pp: 1259-1275, Jan 2022.

[10] H.Park, J.Yoo, G.Venkatesh, N.Kwak, "**Adaptive Template and Transition Map for Real-Time Video Object Segmentation**", IEEE Access, Vol.9, pp: 116914-116926, Aug 2021.

[11] W.Choi, Y.Jun Koh, Chang-Su Kim, **“Video Frame Interpolation Based on Symmetric and Asymmetric Motions”**, IEEE Access, Vol. 11, pp: 22394-22403, Feb. 2023.

[12] J. Usón, J. Cabrera, D. Corregidor and N. García, "**Analysing Foreground Segmentation in Deep Learning Based Depth Estimation on Free-Viewpoint Video Systems"**, 2022 IEEE 12th International Conference on Consumer Electronics (ICCE-Berlin), Berlin, Germany, September 2022.

[13] J.Park, C.Lee, Chang-Su Kim, **“Asymmetric Bilateral Motion Estimation for Video Frame Interpolation”**, arXiv, pp:1-10, Aug 2021.

[14] Kai Xu, Longyin Wen, **“Self-Supervised Deep TripleNet for Video Object Segmentation”**, IEEE Transactions on Multimedia, pp: 1-11, September 2020.

[15] Caelles, Sergi, et al., "**The 2019 davis challenge on vos: Unsupervised multi-object segmentation.**", arXiv preprint arXiv:1905.00737 May 2019.

Questions?

Thank You