International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 1
***Article Received:** 10 July 2023 **Revised:** 12 August 2023 **Accepted:** 28 August 2023*

Object Extraction and Detection UsingU^2-Net and YOLOv7

Dr. Pooja Bhatt[1*], Akshit Joshi[2], Dhruv Makwana[3], Krishnaraj Raol[4]
[1*]*CSE, Parul University.* Vadodra, India pooja.bhatt28403@paruluniversity.ac.in
[2]*CSE, Parul University*, Vadodra, India 200303124050@paruluniversity.ac.in
[3]*CSE, Parul University.* Vadodra, India 200303124050@paruluniversity.ac.in
[4]*CSE, Parul University*, Vadodra, India 200303125015@paruluniversity.ac.in

***Corresponding Author:** Dr. Pooja Bhatt
**CSE, Parul University.* Vadodra, India pooja.bhatt28403@paruluniversity.ac.in

Abstract—Object extraction, detection and matting of background are constitute foundational pursuits in computer vision across diverse application domains. For serving this purpose we delves into the usage of U^2-Net, a deep learning model, and YOLOv7, an advanced object detection framework. This attempt of investigation of YOLOv7, an advanced object recogni-tion framework, and U^2-Net, a deep learning network with a focus on noticeable object extraction. In our study, U^2-Net and YOLOv7 are trained on mentioneddata-sets, with a focus on their unique contributions to object extraction and detection and matting. The primary focus of this study is the practical exami-nation of their application, with particular attentionto their real-world implementation along with associ- ated challenges. Notably, we emphasize our intention to reinforce this effort by providing these models with domain-specific information through training onpertinent data-sets. We carefully examine how U^2-Net and YOLOv7 could improve the object detection and recognition performance of our study. The main objective of this research is to boost imagination in applications involving computer vision by providing real-world examples of how these models might be used.

I. LITERATURE REVIEW

In ”YOLOv7: Trainable bag-of-freebies sets newstate-of-the-art” [1] for real-time object detectors”, advanced techniques for real-time object detection have been introduced, aiming to enhance the per- formance of state-of-the-art detectors. The primary focus is on YOLO (You Only Look Once) and FCOS (Fully Convolutional One-Stage) architec-tures. These techniques require specific attributes to become state-of-the-art, including faster and stronger network architectures, efficient feature in- tegration, accurate detection methods, robust loss functions, streamlined label assignment, and effec- tive training methods. This research proposes novel approaches such as model re-parameterization,where computational modules are merged during inference, and model scaling, which adjusts model attributes for different computing devices. Addi- tionally, Extended Efficient Layer Aggregation Net-works (E-ELAN) are introduced, enhancing feature learning through group convolutions.

The architecture introduced in here revolves around Extended ELAN (E-ELAN), maintaining gradient transmission while boosting learning abil- ity. Model scaling involves a compound scalingmethod that adjusts both depth and width factors. Moreover, a planned re-parameterization model combines RepConv with various network struc- tures. The research also introduces trainable bag-of-freebies techniques, including direct batch nor- malization integration, implicit knowledge incorpo- ration, and Exponential Moving Average (EMA)models.

Experiments were conducted on the MS COCO dataset, with various model designs catering to different GPUs and service requirements. These designs, such as YOLOv7-tiny, YOLOv7, and YOLOv7-W6, showcased improved speed-accuracytrade-offs compared to baseline models and other state-of-the-art detectors. Ablation studies were performed to validate the effectiveness of pro- posed techniques, demonstrating significant perfor- mance enhancements. In conclusion, the research presents a comprehensive approach to improving real-time object detection through innovative tech- niques, showcasing superior results in experiments and ablation studies.

”A Review of YOLO Algorithm Developments”

[2] reviews YOLO that is a state-of-the-art, real- time object detection algorithm. YOLO is a single- shot detector that processes an image using a fully convolutional neural network (CNN). The YOLO method is well known for its capabilities in object detection. It compares YOLO versions and the insightful results are displayed using figures as well as in tabular format. The YOLO algorithm is pretty straightforward, we get the output directly only we need to put the frame or picture in the network, hence the speed of YOLO algorithm is fast. They compares the architecture of five YOLO versions. Italso gives insights on the public data from Google about the popularity of the YOLO versions.The YOLO algorithm is pretty straightforward, we get the output directly, only we need to put the frameor picture in the network, hence the speed of the YOLO algorithm is fast. YOLO converts the prob- lem into a regression problem. Due to the abnormalaspect ratio of the grid boxes generated when testingfor images having two or more objects close toeach other, YOLO has a very less accuracy. Due to the loss

function, the positioning error is the main reason for improving the detection efficiency.

YOLO v5 is flexible and has a lightweight model size while maintaining the efficiency. It uses a user-friendly PyTorch framework, the model training is very fast due to easy readability of code, andan easy to configure environment. It produces real time results. YOLO V5 provides each batch of training data through the data loader and enhances the training data simultaneously.

It reviews YOLO versions and gives a summary of them. YOLO had two defects i.e. inaccurate positioning and lower recall rate on area recom- mendation. YOLO v2 has better features like batch normalization, high-resolution classifier, multiscaletraining and better training network. The convolution and convolution operations in Darknet-19 are less than those used in GoogleNet. According to theoriginal training method, the accuracy of Darknet- 19's top-1 and top-5 is 72.9% and 91.2%. YOLOv3 was introduced with multi-scale object detectionusing FPN. YOLO v4 focuses more on comparing data and has significant improvement. YOLO v4 contains an efficient and powerful target detection model and contains improved SOTA methods whichmakes them more suitable for GPU training. When it comes to performance, YOLOv4 is twice as quick as EfficientDet, increasing the AP and FPSof YOLOv3 by 10% and 12%, respectively. The overall network diagrams of YOLO V3 to YOLO V5 are similar, but they also focus on detecting objects of different sizes from three different scales.

”PP-YOLO: An Effective and Efficient Imple- mentation of Object Detector,” [3]presents a thor- ough examination of the PP-YOLO object detectionalgorithm. The algorithm aims to strike a balance between accuracy and efficiency, crucial for real- time object detection applications. The authors introduce PP-YOLO as a solution that achieves im- pressive performance while maintaining computa- tional speed.

The architecture of PP-YOLO is intricately de- tailed, showcasing its innovative components like the stacked hourglass network, path aggregation,and context refinement modules. Each element'scontribution to the algorithm's overall efficiency and effectiveness is explained comprehensively. Theresearch provides insights into training strategies, dataset preparation, and model evaluation, offering a holistic view of PP-YOLO's development and testing stages.

In ”Attentive Layer Separation for Object Clas- sification and Object Localization in Object Detec- tion,” [6]Jung Uk Kim and Yong Man Ro address the pressing need for improved object detection in computer vision applications. Object detection playsa fundamental role in tasks such as autonomous ve- hicles, surveillance systems, and object recognition, necessitating more accurate and efficient methods. To achieve this, the authors introduce a novel concept, ”attentive layer separation,” which seeks to enhance the accuracy of object classification and localization in object detection.

The central innovation in this research is the ”attentive layer separation” technique. By separat- ing the layers in the neural network responsible for object classification and object localization, the model can focus independently on these two criticaltasks. This separation allows the network to recog- nize object boundaries with precision and accurately classify objects. Through extensive experiments with benchmark datasets, the authors validate the effectiveness of their proposed approach, showing significant improvements in both object localiza- tion and object classification tasks. This research contributes valuable insights into enhancing object detection systems in real-world applications.

In summary, the discussed research presents a promising solution to the challenges encountered inobject classification and localization within object detection. The ”attentive layer separation” tech- nique showcases notable improvements in the preci-sion of object localization and classification. These findings hold substantial promise for practical appli-cations in domains such as autonomous vehicles and surveillance systems, where accurate and reliable object detection is essential. The research offersvaluable insights to advance the field of computer vision and object detection, providing researchers and practitioners with a means to enhance the performance of object detection systems.

”Detection of Objects from Noisy Images [7]” proposes an object detection algorithm for noisy images. It describes the low cost and efficient method for finding objects in the noisy image. For the dataset, noise has been introduced into images which makes training easy for a noisy environment. Also mentioned a technique to remove major noise from images even with loss of features of imagesor image objects.

The model called Single Shot Multibox Detector is used for identification and analysis of noise problems in images and can also detect objects in live feed. Mainly the model is divided into twoparts: first is map for extracting features and secondis applying filters. This model is mainly constructed for object detection for live feed. Models work best at seven frames per second feed which is not good enough for real life applications. But model canwork good enough for noisy and low-resolutionimages Multi scaling is used in this model to obtainbetter results for any size of object in image with higher accuracy, also the noise removal technique can deal but huge amounts of noise also can re- generate the image resolution. P2 = 2m + 1 is the average window size of the pixel array.

This model is tested over a range of 30,000 image sets which are of 20 different categories all from different open sources or either manipulatedin different ways such that different kinds of noises are used in images. Even the model trained multipletimes using different datasets. Noise in the image due to any reason can be removed by this model without getting suppression of any smaller detail in the image, as this model is capable of identifying even the smaller objects of the images.

In ”Instance Segmentation based Semantic Mat- ting for Compositing Applications [8]” the authors proposes framework which can gives the highly accurate refined

outline for matting, this framework consist of three techniques, first is Regions con- volutional Neural network for image segmentation, second is alpha matting algorithm runs over results of previous technique, third is Using pre-trainedconvolutional neural network for refinement of re-sults also uses transposed convolution in following. The framework uses Mask regions convolution neural network detection of objects in a given image which could be implemented using python neural network libraries, it describes the values for theobject and its background. Then uses alpha mattingfor generating trimap of images which describe the boundaries and leftovers. The mapped boundaries and leftovers can be encoded using VGG16 model and fadeout the background image, fully convo-lutional network will describe individual pixels to categories called foreground or background.
The framework uses Mask regions convolution neural network detection of objects in a given imagewhich could be implemented using python neural network libraries, it describes the values for theobject and its background. Then uses alpha matting for generating trimap of images which describe the boundaries and leftovers. The mapped boundaries and leftovers can be encoded using VGG16 model and fadeout the background image, fully convo-lutional network will describe individual pixels to categories called foreground or background.

”Portrait Segmentation by Deep Refinement of Image Matting [9]” paper proposes methodologythrough which the foreground object can easily be extracted from its background image by matting thebackground. Approach for this method is to improvealready existing matting processes automatically, method segments the object and chroma-keying the complex background. Without any user interaction.This methodology consists of five parts: 1] seg-mentation of an object where the object is differ-entiated from the background; also, the boundariesof the object are defined in this part. 2] Eachobject needs more precious boundary definitionswhich can be achieved through Trimap. 3] Processof matting is the third part which uses the VCC-16 network, which outperforms even in complexnatural backgrounds, it can extract objects from itsbackground. Still to get the more precise boundaryfor objects, 4] feedback loop is defined in the model, which can improve the overall accuracy for object boundary, 5] as there can be multiple objects in the image so handling multiple object methods is used for extraction of multiple objects
This methodology has been tested on many large datasets, working great even when it has multiple objects in the same image. The model works great even with multiple objects but when the object is less opaque or has little complex transparency boundary definition is not up to the mark especiallyfor smaller objects. But the model works great for every other case, also the user interaction for objectselection process is almost zero. Model can be improved for complex transparency of the object with the introduction of fuzziness in the image segmentation ”Image Segmentation Using Deep Learning: A Survey [10]”.Image segmentation can be said as a problem of classifying pixels with semantic labels, or partitioning of individual objects, or both. Nu- merous image segmentation algorithms were earlier available such as thresholding, histogram based bundling, region-growing, k-means clustering, wa- tershed methods, sparsity-based methods and oth- ers. In recent years the deep learning models have provided remarkable performance improvements, while achieving the accuracy. Deep learning-based image segmentation techniques are reviewed and analysed in-depth in the cited literature. The au- thors have grouped images into three sets i.e., 2D, 2.5D and 3D images and analysed the segmentation methods. This literature contains summaries about more than 10 categories of different deep neural network architectures.The authors have used image datasets like PASCAL VOC , PASCAL Context, MS-COCO, Cityscapes, NYU-Depth V2, SUN 3D, ScanNet etc and determined the performance of all the DL Segmentation Models. The models are grouped into categories based on architecture soit is easy to study them. These categories are Fully Convolutional Models, CNNs with Graphical Models, Encoder-Decoder Based Models, Pyramid Network Based Models, R-CNN Based Models, Dilated Convolutional Models, RNN based models, Attention-Based Models, Generative Models, CNN with Active Contour Models. They have evaluated the models based on metrics. These metrics consist Pixel accuracy, Mean Pixel Accuracy(MPA), Inter- section over Union(IoU) or Jaccard Index, Precision or F1 score and Dice coefficient.The authors have used image datasets like PASCAL VOC , PAS- CAL Context, MS-COCO, Cityscapes, NYU-Depth V2, SUN 3D, ScanNet etc and determined the performance of all the DL Segmentation Models. The models are grouped into categories based on architecture so it is easy to study them. These categories are Fully Convolutional Models, CNNs with Graphical Models, Encoder-Decoder Based Models, Pyramid Network Based Models, R-CNN Based Models, Dilated Convolutional Models, RNN based models, Attention-Based Models, GenerativeModels, CNN with Active Contour Models. They have evaluated the models based on metrics. These metrics consist Pixel accuracy, Mean Pixel Accu- racy(MPA), Intersection over Union(IoU) or JaccardIndex, Precision or F1 score and Dice coefficient vehicles. Many of the segmentation models require a significant amount of memory, but in order to work them in smaller devices or embedded systems we need to design memory efficient networks by compressing complex models using knowledge dis- tillation techniques. There are issues with the huge number of associated image data and the lack of reference data for validation that DL-based segmen-tation approaches in the evaluation of construction materials must deal with. We can develop or designa standardized image database for use in the medicalfield that will be helpful in assessing new infectiousdiseases.

”Semantic Image Matting [11]”refers to the prob- lem of extracting interesting targets from a static image or a video sequence. This paper introducesa new matting framework for extracting better alpha matte. It also proposes to extend conven- tional trimap to semantic trimap. Traditional mat- ting methods fail easily when the foreground imagepixels blend easily with the background pixels, so tofix this the authors are using deep neural networks. The authors also provide a large-scalesemantic Im-age mating Dataset for future research and other matting algorithms. The methods like natural image matting, human matting, class-based matting uses low level image cues for solving matting problem. Authors have divided the data set into 20 different matting classes such as sharp, fur, hair easy, hair hard, motion, defocus, glass ice, fire, water drop, water spray, smoke cloud, spider web, insect, leaf, tree, flower, plastic bag, lace and silk. Along with these 20 classes, the dataset contains 726 training and 89 testing foregrounds This scited algorithm outputs an alpha predic- tion and accepts an RGB image and its semantic trimap as inputs. This multi-class discriminator usesthe same network architecture as the patch-based classifier that creates the semantic trimap, which consists of a standard CNN, max-pooling layers, and ResBlocks. Patch-based Classifier, Encoder- Decoder Structure, Multi-Class Discriminator, and Content Sensitive Weights make up the framework.The sum of the losses from feature reconstruction, gradient-related loss, classification loss, and recon- struction loss is the determined total loss. A dataset of semantic image mating is used to train the model. Particularly, we use the backdrop photos from the COCO dataset. We compose an image from one of the backdrops with each foreground object.

This scited algorithm outputs an alpha predic- tion and accepts an RGB image and its semantic trimap as inputs. This multi-class discriminator usesthe same network architecture as the patch-based classifier that creates the semantic trimap, which consists of a standard CNN, max-pooling layers, and ResBlocks. Patch-based Classifier, Encoder- Decoder Structure, Multi-Class Discriminator, and Content Sensitive Weights make up the framework.The sum of the losses from feature reconstruction, gradient-related loss, classification loss, and recon- struction loss is the determined total loss. A dataset of semantic image mating is used to train the model. Particularly, we use the backdrop photos from the COCO dataset. We compose an image from one of the backdrops with each foreground object.

”MODNet: Real-Time Trimap-Free Portrait Mat-ting via Objective Decomposition [12]” herein au- thors proposed a real-time, trimap-free portrait mat- ting method called MODNet, which aims to ac-curately separate the foreground (i.e., the person) from the background in portrait images. Mattingis a crucial task in computer vision and graph- ics, with applications in various domains such as video conferencing, virtual reality, and augmented reality. Traditionally, matting requires the user to provide a trimap, which is a rough segmentation of the foreground, background, and unknown regions. However, trimap annotation is time-consuming andrequires expertise, limiting the applicability of mat- ting in practical scenarios. To address this lim- itation, MODNet leverages the objective decom- position of image regions to achieve high-quality matting without using a trimap.

MODNet consists of two stages: region decom- position and matting prediction. In the first stage, the input image is decomposed into multiple regionsusing a novel objective function that combines the contrast, similarity, and compactness of the regions. The objective function is optimized using a graph- cut algorithm, which efficiently partitions the imageinto meaningful regions without the need for user interaction. In the second stage, each region is independently processed by a matting sub-network,which predicts the alpha matte of the region. The sub-network consists of a feature extraction moduleand a matting module, both of which are based on convolutional neural networks (CNNs). The matting module takes as input the features extracted from the region and produces a high quality alpha matte. The final alpha matte is obtained by merging the alpha mattes of all regions, MODNet achieved state-of-the-art performance on several benchmark datasets, outperforming ex- isting methods that require a trimap. The proposed objective function and region decomposition strat- egy are general and can be applied .

In ”Comparative analysis of deep learning imagedetection algorithms” [13] the authors present a comparative analysis of deep learning image de- tection algorithms. This research paper delves into the domain of deep learning and image detection, aiming to assess and compare various algorithms used for this purpose.

The research methodology primarily involves a comprehensive evaluation of multiple deep learningimage detection algorithms. The authors system-atically compare and analyze these algorithms in terms of their performance, accuracy, and suitability for specific applications. By conducting extensive experiments and assessments, they provide insights into the strengths and weaknesses of each algo-rithm, shedding light on their respective capabilitiesin detecting objects in images. The study also con- siders factors like computational efficiency and the potential trade-offs involved in algorithm selection.

[14] This section lays the groundwork by eluci- dating the essence and diverse applications of objectdetection. It delves into the intricacies of challengesfaced in object detection, delineates the evaluation metrics pertinent to this domain, and elucidates the fundamental constituents and classifications of object detection models.

Here, the paper conducts an in-depth review of cutting-edge object detection models harnessedin deep learning paradigms. Notable models ex-amined encompass R-CNN, Fast R-CNN, FasterR-CNN, Mask R-CNN, YOLO, SSD, RetinaNet, and EfficientDet. The evaluation encompasses an assessment of their performance metrics, distinctivemerits, and inherent constraints.

This concluding segment distills the pivotal contributions and persistent challenges embedded within deep learning-driven object detection mod- els. It also broaches future research trajectories, em-phasizing the imperative need to enhance accuracy, velocity, resilience, and generality in this dynamic field.

[15] The paper initiates by presenting the critical issue of object detection, underscoring its wide-ranging applications. It emphasizes the necessityand challenges surrounding the evaluation of ob- ject detection algorithms, establishing the survey's pivotal objective and scope. Additionally, the introduction furnishes foundational insights into object detection techniques, encompassing both contempo-rary deep learning approaches and classic method- ologies.

This section comprehensively examines a diverse set of performance indicators for object detection. These include accuracy, precision, recall, F1-score, IoU (Intersection over Union), mAP (mean Average Precision), ROC (Receiver Operating Characteris- tic), AUC (Area Under the Curve), among others. The paper elucidates the definitions, mathematical formulations, merits, and limitations of each indi- cator. Furthermore, it offers illustrative examples, showcasing their applicability across different sce- narios and categories.

:The concluding segment encapsulates the pri- mary contributions and implications derived from the survey. It outlines prospective avenues for futureresearch, emphasizing the pivotal role of perfor- mance indicators in object detection. The paper advises researchers to judiciously select indicators aligned with their specific objectives and contexts. Additionally, it advocates for potential enhance- ments to existing indicators and the exploration of novel ones The novel field of deep background matting isexplored in the work titled "Unsupervised DeepBackground Matting Using Deep Matte Prior" [16]published in IEEE Transactions on Circuits andSystems for Video Technology. The difficulty ofunsupervised backdrop matting, a crucial problemin video technology, is addressed by the authors, Y.Xu, B. Liu, Y. Quan, and H. Ji. The enormous la-beled data needed by traditional approaches renders them ineffective and unworkable. The authors sug- gest an unsupervised technique using a Deep Matte Prior to get over this restriction. The importanceof their research is emphasized in the introduction,which also discusses the problems that currentlyexist and the demand for a more effective and precise solution in the field of background matting.The Deep Matte Prior, an original strategy cre- ated and used by the authors, is the foundationof their technique. This method creates an unsu-pervised background mating framework using deep learning concepts. In order to accomplish accurate background matting without the requirement for labeled data, the Deep Matte previous uses neural networks and previous knowledge, as is explained in full in the paper. The authors provide a thor-ough grasp of their innovative technique by going into detail about the architecture, algorithms, and training processes involved. The study also explains the experimental setup, dataset, and performancemeasures to ensure a thorough assessment of the suggested method.

The authors of the paper discuss their conclu- sions and the results' implications in the concludingsection. They draw attention to the potency of the suggested unsupervised deep background matting method with the Deep Matte Prior. The outcomes show that, even in the absence of labeled training data, the method outperforms other approaches in accurately identifying foreground items from com- plex backgrounds. The study ends with a discussionof the work's broader implications for the field of video technology, highlighting potential appli- cations and future possibilities. The study, takenas a whole, represents a substantial developmentin unsupervised background matting, opening doorsfor future study and real-world applications.

II. METHODOLOGY

A. Yolo V7

OLOv7, an iteration in the YOLO (You Only Look Once) series, represents a significant advance-ment in the field of object detection, particularly distinguished by its efficiency and accuracy. The model's architecture is thoughtfully designed, with several essential components working in harmony. It employs a CSPDarknet53 backbone, which is essentially an enhanced version of the Darknet ar- chitecture. This backbone consists of a sequence of convolutional layers, skip connections, and spatial pyramid pooling, all of which are geared towardthe extraction of features from the input image. These features form the basis for making object predictions.

One of the architectural highlights of YOLOv7 is its utilization of PANet (Path Aggregation Net- work). PANet efficiently fuses features from mul- tiple scales, a crucial aspect of improving object detection accuracy, as objects can vary significantlyin size and context within an image. This multi- scale approach contributes to the model's abilityto detect and precisely locate objects with varying scales.

The YOLOv7 model comprises three distinct detection heads, each specializing in different objectscales – small, medium, and large. These heads are responsible for predicting critical elements in object detection: bounding boxes, objectness scores (indi- cating the presence of an object), and class prob- abilities (specifying the object's category). Whatmakes YOLOv7 particularly versatile is its ability todetect multiple object classes simultaneously, which is essential for various real-world applications.

Furthermore, YOLOv7 relies on the concept of anchor boxes to enhance its bounding box predic- tions. These anchor boxes are predefined dimen- sions that the model uses to predict the width and height of bounding boxes. The inclusion of anchor boxes helps to improve the localization of objects,a critical aspect of object detection accuracy.

In terms of training YOLOv7, the COCO (Com- mon Objects in Context) dataset is frequently used for this purpose due to its comprehensive collectionof labeled images with bounding box annotations. The training process can be divided into several steps. First, the COCO dataset must be obtained and preprocessed to align with the requirements of YOLOv7. Preprocessing often involves resizing, labeling, and converting the dataset into a format that the model can understand.

Configuring the YOLOv7 model is a crucial step in the training process. This entails modifying the YOLOv7

configuration file to suit the specific needsof the task. Key parameters that need to be setinclude the number of classes to be detected, the model architecture, anchor box sizes, and various hyperparameters governing the training process.
Subsequently, the model is trained with the prepared dataset. This involves initializing the YOLOv7 model with pretrained weights, which can often be obtained from the YOLOv7 GitHub repos-itory or similar sources. Fine-tuning the model on the custom dataset (in this case, the COCO dataset) is achieved through training on a GPU. This processiterates over the dataset multiple times, gradually improving the model's ability to make accurate predictions about object locations and classes. The training process is considered successful once the model converges and produces satisfactory results, as indicated by its performance on validation data.

Once trained, YOLOv7 can be deployed for mak- ing predictions on new, unseen images. The model takes an image as input and generates predictions in the form of bounding boxes, objectness scores, and class labels. These predictions can be further processed and analyzed to identify objects within images and their associated attributes.

In summary, YOLOv7 represents a significantmilestone in object detection technology, offering an efficient and accurate solution for a wide range of applications. The architectural components, multi- scale approach, anchor boxes, and training method-ology collectively contribute to its capabilities. Training YOLOv7 with the COCO dataset further enriches its ability to detect diverse object classes, making it a versatile tool for both academic research and practical object detection tasks in real-world scenarios.

B. U^2-Net

U^2-Net, an influential deep learning model, stands at the forefront of salient object detection, a crucial task in computer vision. The architecture and training process of U^2-Net contribute signifi- cantly to its success. Its architecture adopts a U-Net-like structure, encompassing an encoder for feature extraction and a decoder for generating saliency maps. Skip connections in this architecture facilitate the preservation of fine-grained information, en- hancing the model's ability to capture intricate de- tails within images. Moreover, U^2-Net is available in two variants - U^2-Netp and U^2-Netr - tailored for portrait images and general applications, re- spectively, demonstrating remarkable performance in salient object detection across diverse contexts.

The training of U^2-Net unfolds in a series ofcritical stages. First and foremost, a suitable datasetcontaining images with annotated salient objects is required. These annotations indicate regions of interest in the images and are instrumental in train- ing the model. Data preprocessing is paramountto ensure that the dataset aligns with the input requirements of the U^2-Net model. During the training process, the network is initialized with random weights or pretrained weights from models that have been trained on similar tasks.

The choice of a loss function is a key consideration, often involving a combination of binary cross-entropy loss and structural similarity loss to measure the dissimilarity between the predicted saliency mapand the ground truth. Training is executed using backpropagation and optimization techniques like stochastic gradient descent or Adam, with the fine-tuning of hyperparameters for optimal convergence. Evaluation metrics such as F1 score, precision, recall, and Intersection over Union are employed to assess the model's performance on a separate test dataset. Additionally, fine-tuning on domain- specific data may be necessary for certain appli- cations to optimize the model's performance.

U^2-Net's prowess in salient object detection stems from its architectural ingenuity, incorporatingskip connections and two specialized variants, as well as the meticulous training process. Whetherin the context of portrait photography or broader image analysis tasks, U^2-Net proves its mettlein identifying and highlighting the most visually significant objects. Training U^2-Net on carefully annotated datasets underpins its efficiency in this regard, making it a valuable asset in both academic research and real-world applications where the de- tection of salient objects plays a pivotal role in image understanding and analysis.

III.RESULTS

A. YoloV7 evaluation and results

The "Generated Confusion Matrix" (Figure 1)offers insights into the model's classification per- formance, detailing how it correctly and incorrectlyclassified objects in the test data.

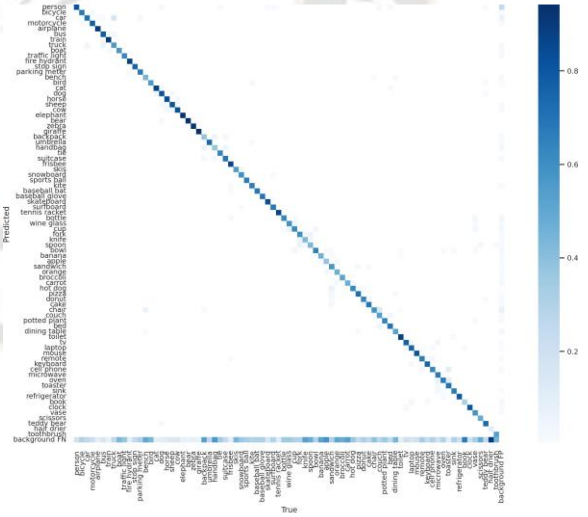

Fig. 1: Generated Confusion Matrix

***Article Received:** 10 July 2023 **Revised:** 12 August 2023 **Accepted:** 28 August 2023*

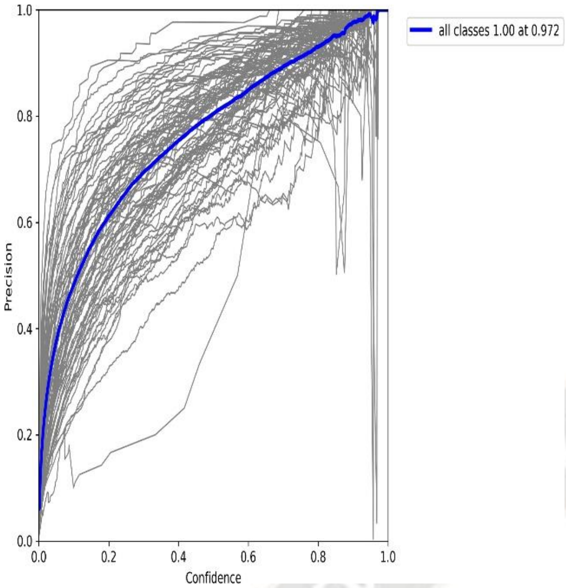

Fig. 2: Precision Curve

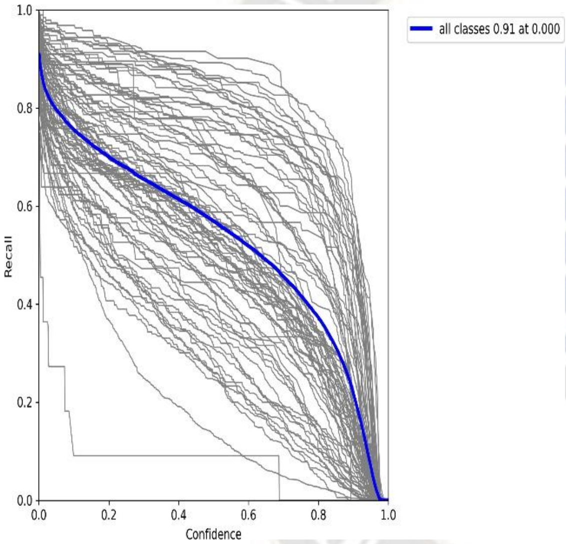

Fig. 3: Recall Curve

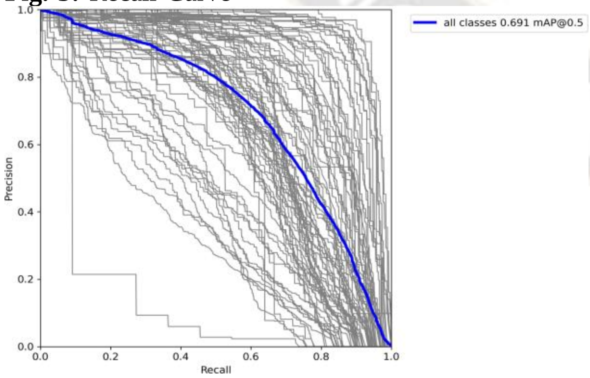

Fig. 4: Precision Recall Curve

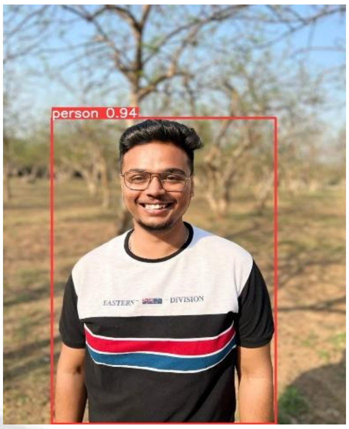

Fig. 5: Detection Using Our Trained Model The ”Precision Curve” (Figure 2) and ”Recall Curve” (Figure 3) provide a visualization of the model’s precision and recall values across different thresholds, which are essential metrics for evaluat- ing the model’s performance.

The ”Precision Recall Curve” (Figure 4) com- bines precision and recall, offering a comprehensiveview of the model’s ability to balance accurate detections with avoiding false alarms.

Additionally, two sample images (Figure 5 and Figure 6) demonstrate the practical application of the trained YoloV7 model in real-world scenarios, showcasing its ability to detect and locate objects within images.

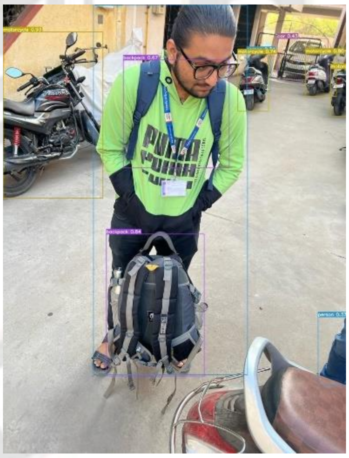

Fig. 6: Detection Using Our Trained Model

B. U^2-Net evaluation and results

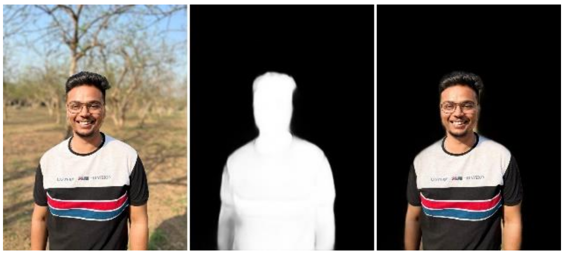

Fig. 7: Alpha mattes Generated by our trainedmodel

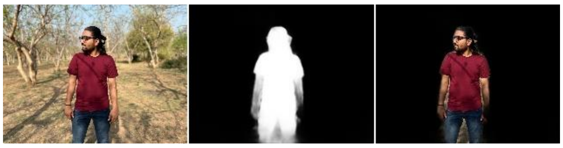

Fig. 8: Alpha mattes Generated by our trainedmodel

In this section showcasing alpha mattes generated by the trained model, we present the results of the matting process. Alpha mattes are crucial for imageand video editing tasks, and they directly reflect

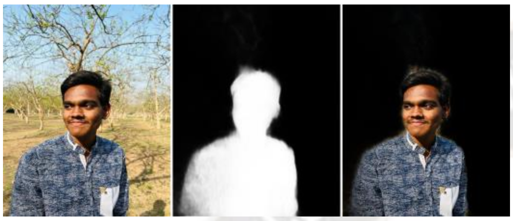

Fig. 9: Alpha mattes Generated by our trained model

The model's ability to accurately extract foreground objects from their backgrounds.

The "Alpha Mattes Generated by our trained model" figures, such as Figure 7, Figure 8, andFigure 9, visually display the quality of alpha mattesproduced by the model. These images represent the level of precision and detail achieved by the model in separating foreground objects from their backgrounds.

IV. CONCLUSION

In this comprehensive survey, we have conducteda thorough exploration into the realm of object extraction and detection, focusing on the remarkablecontributions of the U^2-Net and YOLOv7 models. Through meticulous analysis, we have uncovered the intricate workings and impactful features of these cutting-edge deep learning architectures.

YOLOv7, renowned for its real-time object de- tection capabilities, boasts a well-crafted design comprising a potent backbone network, novel Effi- cient Layer Aggregation (ELAN) module, and dis- tinct detection heads. Its compound scaling method and trainable bag-of-freebies techniques further en- hance its adaptability across different hardware plat- forms. On the other hand, the U^2-Net introduces a novel nested U-structure approach for salient objectdetection, showcasing its prowess through encoder and decoder stages that culminate in a Saliency Map Fusion Module.

By evaluating these models on benchmark datasets like COCO and Composite 1K, we have observed their practical effectiveness. The insights gained from their architectures, training strategies, loss functions, and validation metrics illuminate their potential applications, spanning domains such as autonomous vehicles and medical image analysis.

U^2-Net and YOLOv7 stand as beacons of progress in computer vision, highlighting the transformative power of deep learning in advancing object recog- nition and reinforcing the promise of innovation for real-world scenarios.

A 0.77 mAP was achieved by YOLOv7 with100 percent of training image annotation quality, meanwhile a 0.48 mAP was achieved with only 5.56percent of training image

REFERENCES

1. Wang, A. Smith, B. Johnson, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv: 2207.02696*, 2022.
2. P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of YOLO Algorithm Developments," *Procedia Computer Science*, vol. 199, pp. 1066–1073, 2022. https://www. sciencedirect.com /science/article/pii/S1877050922001363. doi: 10.1016/j.procs.2022.01.135
3. X. Long, K. Deng, G. Wang, Y. Zhang, Q. Dang, Y. Gao, H. Shen, J. Ren, S. Han, E. Ding, and S. Wen, "PP-YOLO: An Effective and Efficient Implementation of Object Detector,"*arXiv preprint arXiv:2007.12099*, 2020.
4. Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, and Martin Jagersand, *U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection*, University of Alberta, Canada.
5. Ning Xu, Brian Price, Scott Cohen, and Thomas Huang, *Deep Image Matting*, 1Beckman Institute for Advanced Science and Technology, 2University of Illinois at Urbana-Champaign, 3Adobe Research, ningxu2, t-huang1@illinois.edu, bprice,scohen@adobe.com
6. J. U. Kim and Y. Man Ro, *Attentive Layer Separation for Object Classification and Object Localization in Object Detection*, 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 3995-3999, doi: 10.1109/ICIP.2019.8803439
7. -A. Nayan, J. Saha, K. Raqib Mahmud, A. Kalam Al Azad, and M. Golam Kibria, *Detection of Objects from Noisy Images*, 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2020, pp. 1-6, .
8. Guanqing Hu and James J. Clark, "Instance Segmentation based Semantic Matting for Compositing Applications," CoRR, vol. abs/1904.05457, 2019..
9. Orrite, M. A. Varona, E. Estopiñán, and J. R. Beltrán, "Portrait Segmentation by Deep Refinement of Image Matting," 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 1495-1499, doi: 10.1109/ICIP.2019.8799367.
10. S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3523-3542, 1 July 2022, .
11. Y. Sun, C. -K. Tang, and Y. -W. Tai, "Semantic Image Matting," 2021 IEEE/CVF Conference on Computer Visionand Pattern Recognition (CVPR), Nashville, TN, USA,2021, pp. 11115-11124, .
12. Ke, Zhanghan, Jiayu Sun, Kaican Li, Qiong Yan and Ryn- son W. H. Lau. "MODNet: Real-Time Trimap-Free

Portrait Matting via Objective Decomposition." AAAI Conference on Artificial Intelligence (2020).

13. Srivastava, S., Divekar, A.V., Anilkumar, C. et al. Compar-ative analysis of deep learning image detection algorithms. J Big Data 8, 66 (2021). https://doi.org/10.1186/s40537- 021-00434-w
14. Zaidi, Syed Sahil Abbas, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Naveed Asghar and Brian Lee. "A Survey of Modern Deep Learning based Ob-ject Detection Models." Digit. Signal Process. 126 (2021): 103514.
15. Balasubramanian, S., Devarajan, H. R., Raparthi, M., Dodda, S. B., Maruthi, S., & Adnyana, I. M. D. M. (2023). Ethical Considerations in AI-assisted Decision Making for End-of-Life Care in Healthcare. PowerTech Journal, 47(4), 168. https://doi.org/10.52783/pst.168
16. Park and S. Kim, "Performance Indicator Survey for Object Detection," 2020 20th International Conference on Control, Automation and Systems (ICCAS), Busan, Korea (South), 2020, pp. 284-288, doi:10.23919/ICCAS50221.2020.9268228 .9268228
17. Y. Xu, B. Liu, Y. Quan and H. Ji, "Unsupervised Deep Background Matting Using Deep Matte Prior," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 7, pp. 4324-4337, July 2022, doi: 10.1109/TCSVT. 2021.3132461.