

This work was written as part of one of the author's official duties as an Employee of the United States Government and is therefore a work of the United States Government. In accordance with 17 U.S.C. 105, no copyright protection is available for such works under U.S. Law.

Public Domain Mark 1.0

<https://creativecommons.org/publicdomain/mark/1.0/>

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

Vision Transformer-based Real-Time Camouflaged Object Detection System at Edge

Rohan Putatunda

Dept. of Information Systems

University of Maryland Baltimore County
rohanp1@umbc.edu

Md Azim Khan

Dept. of Information Systems

University of Maryland Baltimore County
azimkhan22@umbc.edu

Aryya Gangopadhyay

Dept. of Information Systems

University of Maryland Baltimore County
gangopad@umbc.edu

Jianwu Wang

Dept. of Information Systems

University of Maryland Baltimore County
jianwu@umbc.edu

Carl Busart

DEVCOM Army Research Laboratory

Adelphi, Maryland, United States
carl.e.busart.civ@army.mil

Robert F. Erbacher

DEVCOM Army Research Laboratory

Adelphi, Maryland, United States
robert.f.erbacher.civ@army.mil

Abstract—Camouflaged object detection is a challenging task in computer vision that involves identifying objects that are intentionally or unintentionally hidden in their surrounding environment. Vision Transformer mechanisms play a critical role in improving the performance of deep learning models by focusing on the most relevant features that help object detection under camouflaged conditions. In this paper, we utilized a vision transformer (VT) in two phases, a) By integrating VT with a deep learning architecture for efficient monocular depth map generation for camouflaged objects and b) By embedding VT multiclass object detection model with multimodal feature input (RGB with RGB-D) that increases the visual cues and provides more representational information to the model for performance enhancement. Additionally, we performed an ablation study to understand the role of the vision transformer in camouflaged object detection and incorporated GRAD-CAM on top of the model to visualize the performance improvement achieved by embedding the VT in the model architecture. We deployed the model on resource-constrained edge devices for real-time object detection to realistically test the performance of the trained model.

Index Terms—Camouflaged Object Detection, Multi-Modality, Vision Transformer, GRAD-CAM

I. INTRODUCTION

Automatic target recognition (ATR) under camouflaged conditions is a challenging task involving detecting and identifying objects or targets concealed by their surroundings. Camouflage can take various forms, such as natural or artificial foliage, paint, or texture, making it difficult for ATR systems to identify and localize targets.

Camouflaged object detection (COD) has a wide range of useful applications in various fields, including medical diagnosis (e.g., polyp segmentation [1], and lung infection segmentation [2]), industry (e.g., an product inspection, anomaly detection), agriculture (e.g., locust detection to prevent invasion), security and surveillance (e.g., search-and-rescue mission, pedestrian detection, automatic driving in bad weather), scientific research (e.g., rare species discovery), in the field of art (e.g., recreational art and photo blending). The fact that the camouflage strategy deceives the observer's visual perception

system [3] makes COD a fundamentally difficult task, and removing the ambiguities brought on by the high intrinsic similarity between the target object and the background requires a substantial amount of visual perception knowledge [4].

Previous techniques focus on separating the foreground object from the background using handcrafted low-level features, including texture [5], 3D convexity [6], and motion [7]. The extracted low-level features, however, have limited capability to distinguish between camouflaged and non-camouflaged objects, so the approaches based on them often fail in complex scenes. Although the newly proposed deep learning-based systems [8]–[10] have partially improved performance, there is still much opportunity for developing the robust COD model.

Motivated by previous research, the geometric information provided by the depth map makes the observer more sensitive to the true boundary and thus enables camouflage to be less effective. In our work, we propose multimodal visual and perceptual knowledge about camouflaged objects that can be beneficial for object detection under a contested environment; current research works in the object segmentation domain [11]–[12] have an integrated depth map as another modal along with the RGB image to extract features about the 3D layout of the scene and object shape information. As far as we know, no deep camouflaged object detection model utilizing depth for effective object detection for military or traffic assets. In this paper, we approached the problem of camouflaged object detection with the novel multimodal feature fusion architecture integrated with the vision transformer embedded into the architecture of the model to extract high-level discriminative features for developing a robust camouflaged object detection model. Our work developed a well-represented depth map under camouflaged conditions using a deep learning-based monocular method embedded with a vision transformer that provides more representational information for the object detection model.

Additionally, to give a realistic view of the performance of the trained model, we deploy the models in the resource-constrained edge device to see the deployable performance

measures. To study the contribution of information provided by the vision transformer for camouflage detection, we conducted the ablation study for which we did four tiers of experiments with two different types of datasets; first, datasets pertaining to army assets(soldiers, tanks, and mines) and other related to traffic (cars and trucks). Furthermore, to understand the role of the vision transformer in the camouflaged object detection model(COD), we used the GRAD-CAM on top of the COD model to understand the role of multimodality integrated with the vision transformer in feature enrichment that improves the detection model's performance under the contested environment.

To address the other backset of using deep learning models for camouflage object detection requires a high volume of data for appropriate model training for efficient performance. To address the challenge of data paucity under varying scenarios in the battlefield and traffic, we propose a neural-style transfer approach for synthetic data generation [13], which are treated as camouflaged images for our experiments.

Our main contributions to this paper are as follows:

1. We are able to generate a high representational depth map by combining a vision transformer on top of the conventional monocular deep learning-based depth map methodology.
2. We introduced a multimodal feature learning with a state-of-the-art YOLOv5 model and integrated the vision transformer into the architecture of the model to improve the performance of our camouflaged object detection (COD) model.

The rest of the paper is organized as follows. Section. 2 presents the related work of COD, Section.3 describes the methodology in this paper, while Section.4, we discuss the experimental results with model deployment and performance at edge devices.

II. RELATED WORK

The related section in our paper is discussed in three subsections: A) Camouflaged Object Detection, B) The Role of Multimodality in Object Detection, and C) The Significance of the Attention layer in Object Detection.

A. Camouflaged Object Detection

The camouflaged object detection task has posed a new challenge for the current state-of-the-art object detection models. Research in COD [14]–[19] transitioned from conventional methodologies to deep neural networks for extracting high-level semantic features to discriminate the concealed object from the complex scenario. Wang et al. [20] introduce the work of detecting camouflage military tanks with the YOLOv2 deep learning model. In [18] Search, and Fan introduced identification networks (SINet) to address the COD challenge by firstly searching for camouflaged objects and then performing the segmentation. In [16], Classification information was introduced into representational learning in Anabranh Network (ANet) for camouflaged object detection. Yan et al. In [17], introduce the concept of MirrorNet to use instance segmentation and adversarial attack for COD. In [21], the Positioning and focus network (PFNet) discusses the distraction

mining strategy by distraction discovery and removal to benefit estimation performance.

B. The Role of Multimodality in Object Detection

For object detection tasks, many studies have shown that depth information can reduce the ambiguity of RGB features in complex scenarios. Zhang et al. [22] introduce the concept of depth-guided camouflage for the salient object detection task. In recent years, many studies have benefited from powerful deep learning, demonstrating that deep cross-modal fusion for segmentation has improved performance. Farabet et al. [23] propose an early fusion to concatenate the RGB and depth channels and feed the input into a multiscale CNN network. [24] fuses the maps from the depth, RGB, and early fusion branch at the decision level, which is called late fusion. The major challenge in COD is to fuse the high-quality depth maps for the obfuscated RGB image and achieve the same effective segmentation methodologies that suggest the fusion of the complementary information in the middle level of the network [25]–[29].

C. The Significance of the Attention mechanism in Object Detection

The attention layer plays a crucial role in the feature enrichment for the object detection model. Jie et al. proposed a squeeze-and-excitation (SE) module [30]. It is a light plug-in module that allows the network to perform feature recalibration through which the network learns to use global information to emphasize informative features and nullify less useful ones selectively. In their work, global information is obtained by a global average pooling operation. Sanghyun et al. proposed the convolutional block attention module (CBAM) [31]. They gather global information through both global average pooling and global max pooling because global max pooling gathers finer channel-wise attention. Moreover, they also devised a spatial attention module through an inter-spatial relationship of features. Different from channel-wise attention, which focuses on 'what' to attend to, whereas spatial attention focuses on where is an informative part. CBAM gathers channel-wise attention and spatial-wise attention sequentially. Jongchan et al. also exploited both channel and spatial attention and proposed the bottleneck attention module (BAM) [32]. In their work, BAM gathers channel and spatial attention in parallel.

In our research work for camouflage object detection, we utilize the architecture of a generic object detection model (Yolov5) with enhanced multimodal feature input (VT encoder-decoder network for well-represented RGB-D image generation) embedded with the vision transformer on the architecture of the model for higher-level discriminative feature learning to reduce the impact of camouflage in object detection.

III. METHODOLOGY

The methodological section of our paper discusses the synthetic data generation, depth map and improvised depth

map generation, the significance of multimodal features, and improved COD with vision transformer. The overall methodological approach is highlighted in Fig.1.

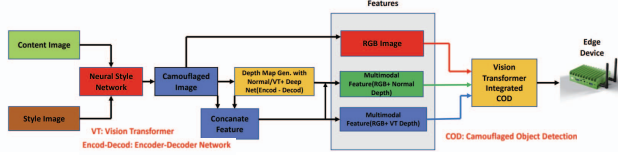


Fig. 1. Overall Methodological Architecture

A. Synthetic Data Generation

Deep learning networks are considered data-intensive models; for appropriate training of those models, we require an adequate amount of data. In our research process, we found that there exist no appropriate camouflage datasets for object detection related to the army or traffic assets. So, we propose the neural style transfer algorithm to generate the appropriate datasets pertaining to army assets (soldiers, mines, and tanks). Additionally, to validate the generalizability of our proposed approach, we generate another dataset related to traffic (cars and trucks). Deep neural style transfer is an optimized image generation technique in which the two images are input to the neural network, i.e., content image and a style image that has contours of the content image and texture, color pattern of the style image, and it is achieved by optimizing the loss function. In our case, we collected the content images and style images [army asset detection (jungle, snow, and desert) and traffic object detection (haze, smoky, and rainy)] from the various publicly available data sources. In our work, we approached the neural style transfer [13], [33] with the VGG16 model with 13 Conv. layers. The concept diagram for the same is shown in Fig.2, in which the content image loss function(mean square error) is optimized by minimizing the difference of the features activated for the content image corresponding to the mixed image, thus preserving the contour of the content image to the resultant mixed image. In contrast, the optimization in the case of the style image is achieved by using the gram matrix, which identifies the gradient of the combined loss function that is updated accordingly to achieve the desired images and referred to as the activated features at a given layer which are considered as the camouflaged images for our experiments.

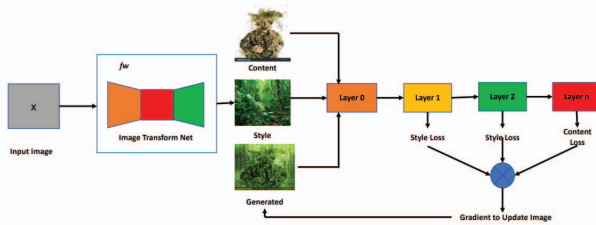


Fig. 2. Concept Diagram for Synthetic Data Generation

B. Depth Map Generation

In our work, to reduce the impact of camouflage objects, we propose to use the depth map as a second feature apart from the RGB image for generating the multimodal feature representation. A depth map provides information about the distance between the surfaces of scene objects from a viewpoint and estimates the geometric relation within a scene. In our research, we approach generating the depth map by utilizing two approaches a) conventional encoder-decoder deep neural network and b) vision transformer integrated encoder-decoder network to generate a well-represented RGB-D feature map under camouflaged conditions. For the first approach, we use DenseNet-169 [34]. An encoder-decoder network using DenseNet-169 for depth map generation is a deep learning architecture that consists of two main parts: an encoder and a decoder. The encoder is a pre-trained DenseNet-169 network that takes an RGB image as input and extracts high-level features from it. DenseNet-169 is a convolutional neural network architecture that is known for its ability to learn features from images efficiently by densely connecting all layers to each other. The pre-trained weights of the DenseNet-169 are used to extract useful features from the input image that are relevant to the depth map generation task. On the other hand, the decoder takes the features extracted by the encoder and processes them to produce the depth map. The decoder typically consists of several upsampling layers with skip connections that gradually increase the spatial resolution of the features. We passed the input RGB image through the DenseNet-169 encoder to extract features. The output of the encoder is then passed through the decoder, which consists of several convolutional and upsampling layers. The output of each convolutional layer is passed through a non-linear activation function ReLU to introduce non-linearity. The final layer of the decoder is a single-channel convolutional layer that produces the depth map as output. The mean-squared error loss function optimizes the network.

In our research, we modified the DenseNet-169 with a vision transformer encoder with a decoder network using DenseNet-169 for depth map generation that combines the strengths of both transformer-based models and convolutional neural networks. The architecture can be divided into two main parts: the encoder and the decoder. The encoder is a vision transformer network that inputs an RGB image and extracts high-level features using self-attention and feedforward neural networks. The transformer architecture is known for its ability to capture global dependencies and relationships between different parts of an image, which makes it suitable for capturing contextual information required for depth map generation. The encoder output is a set of feature maps containing spatial information about the image. The decoder is a convolutional neural network that takes the feature maps generated by the encoder and processes them to produce the final depth map. The decoder consists of several convolutional layers that gradually increase the spatial resolution of the feature maps. The output of each convolutional layer is passed through a

non-linear activation function ReLU to introduce non-linearity. We used the mean-squared error loss function to optimize the network. The overall methodology for depth map/RGB-D is highlighted in Fig.3.

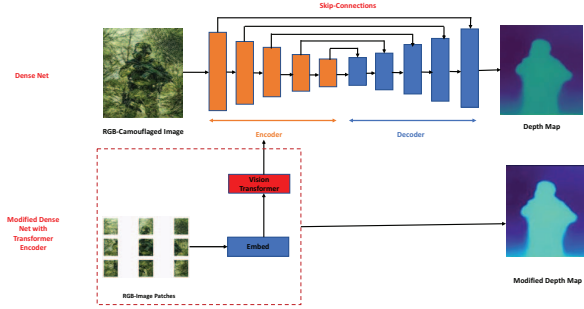


Fig. 3. Methodology for Depth Map Generation

C. Formulation and Significance of Multi-Modal Feature Input

To generate a multimodal feature input that is highlighted in Fig.4, we concatenate camouflaged RGB images with RGB-D images with similar dimensions for better representational learning to reduce the camouflage impacts. Additionally, multimodal systems [35] have several advantages over unimodal systems, in the case of camouflage object detection, which is discussed below: a) Camouflaged images captured in obfuscation, due to texture similarity between foreground object and background scene, are improved with depth maps as a second representational feature, providing guidance and evidence of an object for better model performance. b) Multimodal systems are more reliable because of multiple and independent feature signatures, which reduces the model's error rate. c) In the camouflage object detection problem, the multimodal systems tackle the non-universality problem due to the high noise present in the RGB images, depth map being another feature that tackles the non-universality problem and ensures the process of identification.

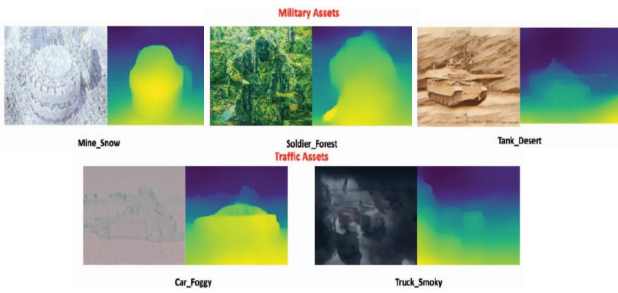


Fig. 4. Multimodal Feature Maps

D. Camouflaged Object Detection Model

In our work for camouflage object detection for the army and traffic asset detection, we present the COD-improved

multimodal system architecture based on the Yolov5. The COD model has the advantage of high detection, accuracy, fast detection speed, and lightweight characteristics, which are suitable for deploying the model in edge devices. We input the multimodal feature into the COD network architecture in our research. The COD consists of three essential blocks: the backbone, neck, and head. The major operational unit blocks present in the input terminal are mosaic data augmentation, adaptive anchor, and image scaling. The Backbone network is a CNN (Convolutional Neural Network), where the image features are extracted from the different aggregated images. The backbone network majorly consists of the focus module, CONV-BN-Leaky ReLU (CBL) module, and CSP1X module with the addition of the vision transformer. Additionally, we also incorporate three transformer layers in the neck section that takes in the feature maps from the backbone where another vision transformer is attached and applies a self-attention mechanism to refine the features. This can help the model better focus on the relevant parts of the image and make more accurate predictions; the feature aggregation layers of mixed and combined image features present in the neck network are mainly used to generate path aggregation networks and feature pyramid networks. The major components of the neck network are the CBL module, Upsample module, CSP2X module, and the vision transformers.

The major advantage of integrating multimodal with vision transformer in COD is to learn better latent space representational features consisting of more information that helps CSPNet architecture to solve the problems of repeated gradient information in large-scale backbones and integrating the gradient changes into the feature map, thereby decreasing the parameters and FLOPS (floating-point operations per second) of the model, which not only ensures the detection speed and accuracy but also reduces the model size.

A feature pyramid network (FPN) is adopted in the PANet to enhance the bottom-up path, improving low-level feature propagation. At the same time, adaptive feature pooling, which links the feature grid and all feature levels, is used to propagate useful information in each feature level directly to the following subnetwork, thus improving the accurate localization of signals in lower layers, which enhances the localization accuracy of the object.

The detection layer present in the output head of the COD model generates varying sizes of feature maps to achieve multi-scale prediction, enabling the model to adapt the various geometrical shape of the object.

The model optimization is done on the summation of the three combined loss functions: classification loss, objectness loss, and regression loss.

$$Total\ Loss = L_{GIoU} + L_{conf} + L_{class} \quad (1)$$

Our COD model with algorithmic approaches is highlighted in Fig.5 and Algorithm.1

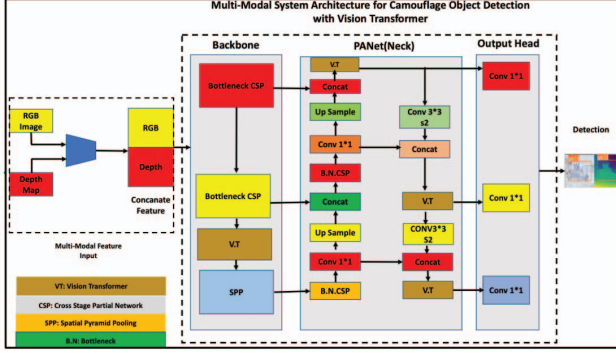


Fig. 5. Camouflaged Object Detection Model

Algorithm 1: Algorithm for Our Camouflaged Object Detection Model

Input: A set of Camouflaged Images;
Output: Trained COD Model;
Preprocess the dimension of an Image 640*640;
Perform the grid cells on top of the Image 16*16;
Perform the bounding boxes on each cell;
Perform each bounding boxes coordinate for box center height, width, and confidence;
Predict the probability of each object class for each bounding box;
Apply non-max suppression to remove overlapping bounding boxes with low confidence scores ;
Initialize COD DNN;
Input feature set to DNN;
while $accuracy \leq 70\%$ **or** $\# epochs \leq 50$ **do**
 Improvise Multimodal Feature Input with depth gen.by VT-Encod Decod ;
 Integrate Vision Transformer in the architecture of COD;
 Compute Total Loss backprop;
end

IV. EXPERIMENTAL SETUP

In this section of the paper, we will discuss the two different data scenarios used for camouflaged object detection and validate the improvement in performance achieved by an improvised multimodal feature input system integrated with the vision transformer in the architecture of the camouflaged object detection model.

A. Dataset

We train our camouflaged object detection model under two different data scenarios, which are broadly discussed below:

First Scenario: The first scenario dataset is related to army asset detection, represented by three classes of soldiers, tanks,

and mines. The total volume of the dataset used for these experiments is 4832 RGB images, in which the soldier's class is composed of 2832 data points and 1000 data points each for the tanks and mines. The three basic style images (jungle/forest, desert, and snow) are used with 55 variations to generate the camouflaged images for the experiments. In addition to RGB camouflaged images, we generate 4832 depth maps, each using two different methodologies totaling 9664 depth map data points (encoder-decoder monocular depth map and vision integrated encoder with decoder) and concatenate feature maps with a similar distribution to the RGB image data points.

Second Scenario: To validate the improved performance achieved by the multimodal architecture, we generate another dataset related to traffic asset detection, which is composed of two classes, i.e., cars and trucks. The total volume of data used for these experiments is 2000, where both classes are represented equally. In this case, we also used three different style images (haze/foggy, smoky, and rainy) with 28 variations to generate the camouflaged images for the improved performance validation experiments. In similar lines to the first scenario, over here, we generate 2000 depth maps utilizing two different (total depth map is 4000) methodologies as mentioned in the first scenario and concatenate feature maps with similar distribution to the RGB image data points.

B. Model Training and Validation Parameters

We trained our camouflaged object detection model with 50 epochs. The model currently has 214 layers with 7.2 million training parameters, 7.2 million gradients, 16.4 GFlops, and we used the batch size of 25 images during training.

Additionally, in this paper, we evaluate the model performance with the various parameters of object detection. The validation parameters like precision, recall, F1 score, and mean average precision (mAP) are calculated as per the confusion matrix, which are described below:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 score} = \frac{TP}{TP + .5(FP + FN)} \quad (4)$$

$$\text{mAP} = \frac{1}{c} \sum_{K=i}^N P(K) \Delta R(K) \quad (5)$$

In Equations 2, 3 and 4, TP represents true positive, FP represents false positive, and FN represents false negative, while in Equation 5 for mAP, c represents the number of object categories, N represents the number of IoU thresholds, k is the IoU threshold, P(k) is precision, and R(k) is the recall.

V. RESULTS AND DISCUSSION

In our paper, we conducted four tiers of experiments for both the scenario (army asset detection and traffic asset detection); the first three experiments are on the RGB image, multimodal image (RGB with RGB-D), multimodal image (RGB with improvised RGB-D) followed by improvised multimodal feature integrated with vision transformer in the architecture to improve the performance of the detection. The improvised multimodal feature is achieved by integrating a vision transformer to generate the well-represented depth map. The data is split into a 70:30 ratio between the training and validation set for the above-mentioned four model scenarios. The first three experiments with the RGB and Multimodal systems clearly highlight the low performance mostly on every evaluation parameter, when compared with the improvised multimodal input feature integrated with the vision transformer in the model architecture highlighted in Table I and Table II. We highlighted the result tables in the form of an ablation study to understand the role of the vision transformer for object detection in camouflaged conditions.

TABLE I

Army Asset Detection with different COD mode				
RGB only				
Class	Precision	Recall	mAP@.5	F1-Score
All	0.814	0.718	0.75	0.76
Soldier	0.775	0.738	0.76	0.75
Mine	0.779	0.582	0.63	0.68
Tank	0.887	0.834	0.87	0.86
RGB + Depth				
Class	Precision	Recall	mAP@.5	F1-Score
All	0.855	0.73	0.79	0.78
Soldier	0.796	0.785	0.81	0.79
Mine	0.911	0.528	0.66	0.71
Tank	0.859	0.878	0.89	0.86
Improvised Multimodal feature with VT Depth				
Class	Precision	Recall	mAP@.5	F1-Score
All	0.86	0.74	0.81	0.79
Soldier	0.8	0.79	0.82	0.81
Mine	0.92	0.54	0.67	0.72
Tank	0.87	0.89	0.89	0.88
Our model with VT				
Class	Precision	Recall	mAP@.5	F1-Score
All	0.89	0.76	0.82	0.83
Soldier	0.83	0.82	0.83	0.83
Mine	0.94	0.57	0.69	0.84
Tank	0.89	0.91	0.9	0.9

We did the first three experiments as a part of an ablation study to determine the importance of a vision transformer which is reflected in the fourth experiment under two different data scenarios that helps the COD model to extract and focus on the high-level discriminative features that are obtained from both RGB image and the depth images where the information about the object distance from the scene is captured.

TABLE II

Traffic Asset Detection with different COD mode				
RGB only				
Class	Precision	Recall	mAP@.5	F1-Score
All	0.654	0.686	0.69	0.67
Car	0.665	0.702	0.7	0.68
Truck	0.644	0.671	0.68	0.65
RGB + Depth				
Class	Precision	Recall	mAP@.5	F1-Score
All	0.781	0.645	0.73	0.71
Car	0.861	0.58	0.75	0.72
Truck	0.702	0.71	0.72	0.7
Improvised Multimodal feature with VT Depth				
Class	Precision	Recall	mAP@.5	F1-Score
All	0.793	0.667	0.75	0.73
Car	0.873	0.59	0.78	0.73
Truck	0.714	0.723	0.74	0.71
Our model with VT				
Class	Precision	Recall	mAP@.5	F1-Score
All	0.816	0.686	0.77	0.751
Car	0.896	0.609	0.8	0.752
Truck	0.737	0.742	0.76	0.741

Additionally, from the model performance perspective, we also observe that total loss for the vision transformer integrated with the COD model is minimized when compared to the three previous systems for both data scenarios, which clearly highlights the better feature space learning for the vision transformer integrated improvised multimodal systems in comparison to the others, the results of which are highlighted in Fig.6.

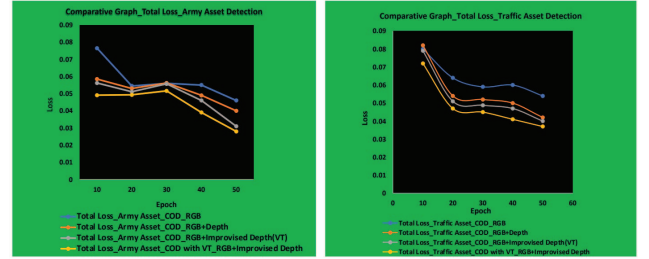


Fig. 6. Total Loss on COD Models

A. Grad-CAM Visualization on COD Model

In our research, we applied the Grad-CAM methodology to understand the importance of a vision transformer integrated with multimodal feature input for the COD model in extracting a very high-level discriminative feature for decision-making comparison to other methodologies. Grad-CAM, also known as (Gradient-weighted Class Activation Mapping) is a technique used for visualizing and understanding the regions of an image that are important for a deep neural network's prediction. The COD model uses a fully convolutional network

to directly predict bounding boxes and class probabilities from raw image pixels. Our COD model's final convolutional layer is the backbone network's last layer before the prediction head. We compute the gradients of the predicted class probability with respect to the feature maps of the final convolutional layer and the average gradients across each channel of the feature maps. This gives us an essential weight for each channel of the final convolutional layer. We then use these weights to compute a weighted sum of the feature maps, generating a heatmap highlighting the image regions that were most important for the prediction. To generate the final visualization, we overlay the heatmap on top of the original RGB test image, using a color map to indicate the strength of the activations. This allows us to visualize the regions of the image that were most important for the COD prediction for a given class, which can be useful for understanding how the model is making its decisions and identifying potential areas for improvement. In Fig.7. above, we can observe that our

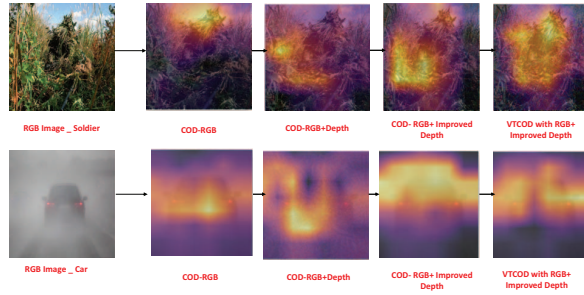


Fig. 7. Grad-CAM on Different Object Detection Models

COD model, with the integration of the vision transformer, accurately captures the core features for making decisions compared to other approaches/models.

B. Real-Time Camouflaged Object Detection System at Edge

Hardware Setup and Performance Evaluation: To test the model in realistic conditions, we deployed the COD algorithm (14MB model weight) on an edge device for real-time inferencing with daily life operational data. We used NVIDIA Jetson Nano, a small and cost-effective single-board computer with high processing power and energy efficiency, to perform complex deep learning and artificial intelligence workloads. It has a quad-core ARM Cortex-A57 CPU, a 128-core NVIDIA Maxwell GPU, and 4GB of LPDDR4 memory. Additionally, it features various I/O interfaces such as USB 3.0, Gigabit Ethernet, HDMI, DisplayPort, and MIPI CSI-2 camera connectors. It runs on Linux-based NVIDIA JetPack SDK, which includes the CUDA toolkit.

To assess the edge suitability of our model and its usability in the camouflage, we migrate model code to NVIDIA Jetson Nano 4GB to use as an edge processor and Intel RealSense Camera as a sensor device. We have written a python script to control the brightness, contrast, exposure, gain, and gamma

parameters of the real sense camera so that our proposed camouflage model can be tested in low-light conditions. Edge is also connected to Orbi WiFi channels (802.11g) for the testing of ground truth. The real-time system setup for COD is highlighted below in Fig.8.

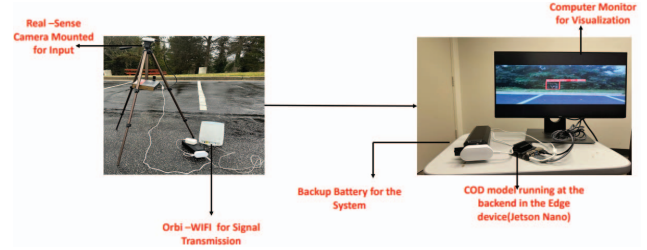


Fig. 8. Real-Time COD at Edge

Our model's inference speed and power efficiency were assessed on a Jetson Nano-embedded module across varying power budgets. Specifically, the model achieved inference speeds of 25 FPS and 15 FPS at 10W and 5W power budgets, respectively, which corresponded to power efficiencies of 2.5 images/sec/watt and 3 images/sec/watt, respectively.

Each image has a large size of 640x640, which can be processed in near real-time. However, during peak usage, the 4GB memory on Jetson Nano was fully utilized, and the system had to use 1.828 GB of swap space on the disk to compensate for the memory shortage. This resulted in substantial throttling of the performance due to disk I/O. We anticipate a significantly better performance on Jetson TX2, as it has more memory and would rarely need to use swap space. This improvement in memory would enable faster inference and bring the processing closer to real-time.

VI. CONCLUSION

In this paper, we tackle the problem of camouflaged object detection (COD) with the enhanced multimodal feature input where the vision transformer encoder-decoder network has generated the well-represented monocular depth map. Furthermore, we also incorporate a vision transformer in the COD model in various stages to enhance the performance of the object detection model by extracting high-level core discriminative features that are essential for object detection under camouflage or obfuscated scenarios (objects concealed under the environment). We validate the model's performance by utilizing various validation parameters and GRAD-CAM methodology. Due to the efficient performance and lightweight nature of the trained model, we are able to successfully deploy the model in a resource-constrained edge device integrated with input-output systems with a quick and precise inference time for object detection under camouflaged conditions.

VII. FUTURE WORK

In the future, firstly, we would like to focus on small-scale camouflage object detection systems, which can be utilized for real-time wide-area surveillance. Secondly, we

would like to contribute more to synthetic data generation for the camouflaged or partially obfuscated objects that can simulate real-time capability. Thirdly, we would implement the channel pruning methodology to optimize model parameters for efficient performance at the edge level.

VIII. ACKNOWLEDGMENT

This work is supported by U.S. Army Grant No: W911NF2120076

REFERENCES

- [1] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, pages 263–273. Springer, 2020.
- [2] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637, 2020.
- [3] Martin Stevens and Sami Merilaita. Animal camouflage: current issues and new perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1516):423–427, 2009.
- [4] Tom Troschianko, Christopher P Benton, P George Lovell, David J Tolhurst, and Zygmunt Pizlo. Camouflage and visual perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1516):449–461, 2009.
- [5] P Sengottuvelan, Amitabh Wahi, and A Shanmugam. Performance of decamouflaging through exploratory image analysis. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 6–10. IEEE, 2008.
- [6] Yuxin Pan, Yiwang Chen, Qiang Fu, Ping Zhang, Xin Xu, et al. Study on the camouflaged target detection method based on 3d convexity. *Modern Applied Science*, 5(4):152, 2011.
- [7] Jianqin Yin Yanbin Han Wendi Hou and Jinping Li. Detection of the mobile object with camouflage color under dynamic background based on optical flow. *Procedia Engineering*, 15:2201–2205, 2011.
- [8] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabran network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019.
- [9] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020.
- [10] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *Scientia Sinica Informationis*, 6(6), 2021.
- [11] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, pages 275–292. Springer, 2020.
- [12] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8582–8591, 2020.
- [13] Rohan Putatunda, Aryya Gangopadhyay, Robert F Erbacher, and Carl Busart. Camouflaged object detection system at the edge. In *Automatic Target Recognition XXXII*, volume 12096, pages 177–187. SPIE, 2022.
- [14] Yunfei Zheng, Xiongwei Zhang, Feng Wang, Tiejong Cao, Meng Sun, and Xiaobing Wang. Detection of people with camouflage pattern via dense deconvolution network. *IEEE Signal Processing Letters*, 26(1):29–33, 2018.
- [15] Zheng Fang, Xiongwei Zhang, Xiaotong Deng, Tiejong Cao, and Changyan Zheng. Camouflage people detection via strong semantic dilation network. In *Proceedings of the ACM Turing Celebration Conference-China*, pages 1–7, 2019.
- [16] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabran network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019.
- [17] Jinnan Yan, Trung-Nghia Le, Khanh-Duy Nguyen, Minh-Triet Tran, Thanh-Toan Do, and Tam V Nguyen. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300, 2021.
- [18] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020.
- [19] Bo Dong, Mingchen Zhuge, Yongxiong Wang, Hongbo Bi, and Geng Chen. Accurate camouflaged object detection via mixture convolution and interactive fusion. *arXiv preprint arXiv:2101.05687*, 2021.
- [20] Yong Wang, Ling Li, Xin Yang, Xinxin Wang, and Hui Liu. A camouflaged object detection model based on deep learning. In *2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIS)*, pages 150–153. IEEE, 2020.
- [21] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8772–8781, 2021.
- [22] Jing Zhang, Yunqiu Lv, Mochu Xiang, Aixuan Li, Yuchao Dai, and Yiran Zhong. Depth-guided camouflaged object detection. *arXiv preprint arXiv:2106.13217*, 2021.
- [23] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2012.
- [24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [25] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 4980–4989, 2017.
- [26] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5):1239–1285, 2020.
- [27] Di Lin, Guangyong Chen, Daniel Cohen-Or, Pheng-Ann Heng, and Hui Huang. Cascaded feature network for semantic segmentation of rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 1311–1319, 2017.
- [28] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for rgb-d salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2343–2350, 2016.
- [29] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for rgb-d saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13756–13765, 2020.
- [30] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [31] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
- [32] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.
- [33] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.
- [34] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [35] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.