MONOCULAR DEPTH ESTIMATION FOR AUTONOMOUS DEVICES

[1]T.Sai Prasad Reddy, [2]V.Chaitanya

[1]Associate Professor, [2]Assistant Professor, [1,2]Department of Computer Science & Engineering, Geethanjali Institute of Science and Technology, Gangavaram, Andhra Pradesh, India

ABSTRACT

This work presents Depth Anything, a highly practical solution for robust monocular depth estimation. Without pursuing novel technical modules, we aim to build a simple yet powerful foundation model dealing with any images under any circumstances. To this end, we scale up the dataset by designing a data engine to collect and automatically annotate large-scale unlabeled data (~62M), which significantly enlarges the data coverage and thus is able to reduce the generalization error. We investigate two simple yet effective strategies that make data scaling-up promising. First, a more challenging optimization target is created by leveraging data augmentation tools. It compels the model to actively seek extra visual knowledge and acquire robust representations. Second, an auxiliary supervision is developed to enforce the model to inherit rich semantic priors from pre-trained encoders. We evaluate its zero-shot capabilities extensively, including six public datasets and randomly captured photos. It demonstrates impressive generalization ability. Further, through fine-tuning it with metric depth information from NYUv2 and KITTI, new SOTAs are set. Our better depth model also results in a better depth-conditioned ControlNet.

Introduction

Depth anything is an area of computer vision that involves estimating the distance between imaged objects and the camera. It allows for understanding a scene’s three-dimensional structure from two-dimensional data. Using artificial intelligence (AI), depth analysis allows machines to perceive the world more like humans. This empowers them to perform tasks like objectdetection, scene reconstruction, and navigating 3D space. Depth Anything is a groundbreaking approach to monocular depth estimation. It effectively harnesses a combination of 1.5 million labeled images and over 62 million unlabeled images. This is a significant differentiation from traditional techniques, Which primarily relied on smaller, labeled datasets. Leveraging the power of large-scale unlabeled data offers a more robust solution for understanding complex visual scenes. For unlabeled images, the model applies consistency loss. This process encourages the model to produce similar depth predictions for slightly perturbed versions of the same image. Depth Anything to improve identity verification processes. It improves security by enabling the system to better discern between a real person and a photo or video. In AR experiences, its precise depth estimation allows for the easy integration of digital objects into real-world scenes. This could greatly simplify complex scene construction tasks in gaming, education, and retail. For autonomous vehicles, the ability to accurately perceive and understand the 3D structure oftheenvironmentfrommonocularimagescancontributetosafernavigation.

OBJECTIVE

The objective of Depth Anything is to provide a highly practical solution for robust monoculardepth estimation. It aims to build asimpleyet powerful foundation model that can deal with any images under any circumstance The core innovation of Depth Anything lies in harnessing the potential of unlabeled images4. The model generates pseudo labels for these images by passing them through a pre-trained Monocular Depth Estimation (MDE) model, resulting in a pseudo-labeled.

Without pursuing novel technical modules, Depth Anything focuses on building a model that can handle a wide range of images and scenarios3. It leverages the power of large- scaleunlabeled data,which significantly enlarges thedatacoverage and thus is ableto reducethe generalization. In summary, the objective of Depth Anything is to unleash the power of large- scale unlabeled data and provide a robust solution for monocular depth estimation

Literature Review

METER:Amobilevisiontransformerarchitectureformonocular depth estimation
LorenzoPapa,PaoloRussoIreneAmerini

Depth estimation is a fundamental knowledge for autonomous systems that need to assess their own state and perceive the surrounding environment. Deep learning algorithms for depth estimation have gained significant interest in recent years, owing to the potential benefits of this methodologyin overcoming the limitations of active depth sensing systems. Moreover, due to the low cost and size of monocular cameras, researchers have focused their attention on monocular depth estimation (MDE), which consists in estimating a dense depth map from a single RGB video frame. State of the art MDE models typically rely on vision transformers (ViT) architectures that are highly deep and complex, making them unsuitable for fast inference on devices with hardware constraints.

Purposely, in this paper, we address the problem of exploiting ViT in MDE on embedded devices. Those systems are usually characterized by limited memory capabilities and low-power CPU/GPU. We propose METER, a novel lightweight vision transformer architecture capable of achieving state of the art estimations and low latency inference performances on the considered embedded hardwares: NVIDIA Jetson TX1 and NVIDIA Jetson Nano.

AModelZooforRobustMonocularRelativeDepthEstimation Reiner Birkl, Diana Wofk, Matthias Mooller

We release MiDaS v3.11 for monocular depth estimation, offering a variety of new models based on different encoder backbones. This release is motivated by the success of transformers in computer vision, with a large variety of pretrained vision transformers now available. We explore how using the most promising vision transformers as image encoders impacts depth estimation quality and runtime of the MiDaS architecture. Our investigation also includes recent convolutional approaches that achieve comparable quality to vision transformers in image classification tasks. While the previous release MiDaS v3.0 solely leverages the vanilla vision transformer ViT, MiDaS v3.1 offers additional models based on BEiT, Swin, SwinV2, Next-ViT and LeViT. These models offer different performance- runtime tradeoffs. The best model improves the depth estimation quality by 28% while efficient models enable downstream tasks requiring high frame rates. We also describe the general process forintegrating new backbones.\

Monoculardepthestimationreferstothetask of regressingdensedepthsolelyfrom asingle inputimage orcameraview. Solvingthisproblem hasnumerousapplications indownstreamtasks like generativeAI 3D reconstruction and autonomous driving . However, it is particularly challenging to deduce depth information at individual pixels given just a single image, as monocular depth estimation is an under constrained problem. Significant recent progress in depth estimation can be attributed to learning-based methods. In particular, dataset mixing and scale-and-shift-invariant loss construction has enabled robust and generalizable monocular depth estimation with MiDaS .

ZeoDepth:Zero_shotTransferbyCombiningRelativeandMetricDepth

ShariqFarooqBhat,ReinerBirkl,DianaWofk,PeterWonka,MatthiasMuller.

This paper tackles the problem of depth estimation from a single image. Existing work either focuses on generalization performance disregarding metric scale, i.e.relative depth estimation, or state- of-the-art results on specific datasets, i.e.metric depth estimation. We propose the first approach that combines both worlds, leading to a model with excellent generalization performance while maintaining metric scale. Our flagship model, ZoeD-M12-NK, is pre-trained on 12 datasets using relative depth and fine-tuned on two datasets using metric depth. We use a lightweight head with a novel bin adjustment design called metric bins module for each domain. During inference, each input image is automatically routed to the appropriate head using a latent classifier. Our framework admits multiple configurations depending on the datasets used for relative depth pre-training and metric fine-tuning. Without pre- training, we can already significantly improve the state of the art (SOTA) on the NYU Depth v2 indoor dataset. Pre-training on twelve datasets and fine-tuning on the NYU Depth v2 indoor dataset, we can further improve SOTA for a total of 21% in terms of relative absolute error (REL). Finally, ZoeD-M12- NK is the first model that can jointly train on multiple datasets (NYU Depth v2 and KITTI) without a significant drop in performance and achieve unprecedented zero-shot generalization performance to eight unseen datasets from both indoor and outdoor domains.

OnthemetricsforEvaluatingMonocularDepthEstimation.

AkhilGurram,Antonio,M.Lopez

Monocular Depth Estimation (MDE) is performed to produce 3D information that can be used in downstream tasks such as those related to on-board perception for Autonomous Vehicles (AVs) ordriver assistance. Therefore, a relevant arising question is whether the standard metrics for MDE assessment are a good indicator of the accuracy of future MDE-based driving-related perception tasks. We address this question in this paper. In particular, we take the task of 3D object detection on point cloudsasaproxyofon-boardperception. Wetrain andteststate-of-the-art3Dobjectdetectorsusing 3D point clouds coming from MDE models. We confront the ranking of object detection results with the ranking given by the depth estimation metrics of the MDE models. We conclude that, indeed, MDE evaluation metrics give rise to a ranking of methods that reflects relatively well the 3D object detection results we may expect. Among the different metrics, the absolute relative (abs-rel) error seems to be the best for that purpose.

Monocular depth estimation (MDE) is addressed from different settings determined by the dataavailable at training time, e.g., LiDAR and virtual-worldsupervision, stereo [8] and structure-from- motion (-self-supervision, and combinations of those . MDE results are compared by using de facto standard metrics (e.g., absrel, rms, etc.) established by Eigen et al. . Reviewing literature results, we can observe that, in terms of such MDE metrics, the difference among different proposals is not too large even the way of training the model is quite different.

Groundembeddingformonoculardepthestimation

Xiaodong yang , Zhuang maqcraft, Zhiyuji

Monocular depth estimation is an ill-posed problem as the same 2D image can be projected from infinite 3D scenes. Although the leading algorithms in this field have reported significant improvement, they are essentially geared to the particular compound of pictorial observationsand camera parameters (i.e., intrinsics and extrinsics), strongly limiting their generalizability in real-world scenarios. To cope with this challenge, this paper proposes a novel ground embedding module to decouple camera parameters from pictorial cues, thus promoting the generalization capability. Given camera parameters, the proposed module generates the ground depth, which is stacked with the input image and referenced in the final depth prediction. A ground attention is designed in the module to optimally combine ground depth with residual depth. Our ground embedding is highly flexible and lightweight, leading to a plug-in module that is amenable to be integratedintovariousdepthestimationnetworks.Experimentsrevealthatourapproach achieves the state-of-the-art results on popular benchmarks, and more importantly, renders significant generalization improvement on a wide range of cross-domain tests.

Accurate depth acquisition is crucial for many robotics application as depth provides pivotalinformation foronboard tasks ranging from perception , predictionto planningAlthough range sensors (e.g., LiDAR) are widely used to produce precise depth measurements, there has been fast growing attentiontocamerabased depthestimationfrombothacademiaandindustryduetoitsportabilityandA typical monocular depth estimation network adopts an encoder-decoder architecture, which can be trained in a supervised or self-supervised mannerMost of the existing works in this field focus on designing more advanced network architecture or engineering more effective loss function.

Towardszeroshotmetric3Dpredictionfromasingle image

Weiyin,chiZhango,haochen,gangyu

Reconstructing accurate 3D scenes from images is a long-standing vision task. Due to the ill-posedNess of the single-image reconstruction problem, most well-established methods are built upon multi- view geometry. State-of-the-art (SOTA) monocular metric depth estimation methods can only handle a single camera model and are unable to perform mixed-data training due to the metric ambiguity. Meanwhile, SOTA monocular methods trained on large mixed datasets achieve zero-shot generalization by learningaffine-invariant depths, which cannot recoverreal-worldmetrics. In this work, weshowthat the key to a zero-shot single-view metric depth model lies in the combination of large-scale datatraining and resolving the metric ambiguity from various camera models. We propose a canonical camera space transformation module, which explicitly addresses the ambiguity problems and can be effortlessly plugged into existing monocular models. Equipped with our module, monocular models can be stably trained over 8 millions of images with thousands of camera models, resulting in zero-shot generalization to in-the-wild images with unseen camera settings. Experiments demonstrate SOTA performance of our method on 7 zero-shot benchmarks. Ourmethod can recover the metric 3D structure on randomly collected Internet images, enabling plausible single-image metrology. Downstream tasks can also be significantly improved by naively plug-in our model. E.g., our model relieves the scale drift issues of monocular-SLAM , leading to metric scale high-quality dense mapping.

MTFormer:Multi-taskLearningviaTransformerandCross-Task Reasoning

XiaogangXu,HengshuangZhao,VibhavVineet,Ser-NamLim,Antonio

"In this paper, we explore the advantages of utilizing transformer structures for addressing multi- task learning (MTL). Specifically, we demonstrate that models with transformer structures are more appropriate for MTL than convolutional neural networks (CNNs), and we propose a novel transformer-based architecture named MTFormer for MTL. In the framework, multiple tasksshare the same transformer encoder and transformer decoder, and lightweight branches are introduced to harvest task-specific outputs, which increases the MTL performance and reducesthe time-space complexity. Furthermore, information from different task domains can benefiteach other, and we conduct cross-task reasoning. We propose a cross-task attention mechanism for further boosting the MTL results. The cross-task attention mechanism brings little parameters and computations while introducing extra performance improvements. Besides, we design a self- supervised cross-task contrastive learning algorithm for further boosting the MTL performance. Extensive experiments are conducted on two multi-task learning datasets, on which MTFormer achieves state-of-the-art results with limited network parameters and computations. It also demonstrates significant superiorities for few-shot learning and zero-shot learning.

The introduction of "Mtformer: Multi-task learning via transformer and cross-task reasoning" likely provides background information on multi-task learning (MTL) and its importance in various machine learning applications. It might discuss the challenges faced in traditional single-task learningapproaches, such as data inefficiency and difficulty in transferring knowledge across tasks. The introduction might also highlight the potential of transformer-based models in capturing complex patterns in data and their success in various natural language processing and computer vision tasks.

EXISTING SYSTEM

MiDaS, which stands for Multiple Depth Estimation Accuracy with Single Network, is a deep learning model designed for monocular depth estimation. This means it can estimate the depth of objects from a single 2D image. These models offer different performance-runtime trade-offs. The best model improves the depth estimation quality by 28% while efficient models enable downstream tasks requiring high frame rates. The MiDaS architecture was trained on up to 12 datasetswith multi- objective optimization. TheMiDaS is used to compute depth from a single image. It provides a variety of models to choose from, depending on the quality and speed-performance trade-off that suits your needs.

SYSTEMARCHITECTURE

The System consists of the following steps :-

1. Input Image
2. Image Encoder
3. Depth Estimation
4. Decoder

Output

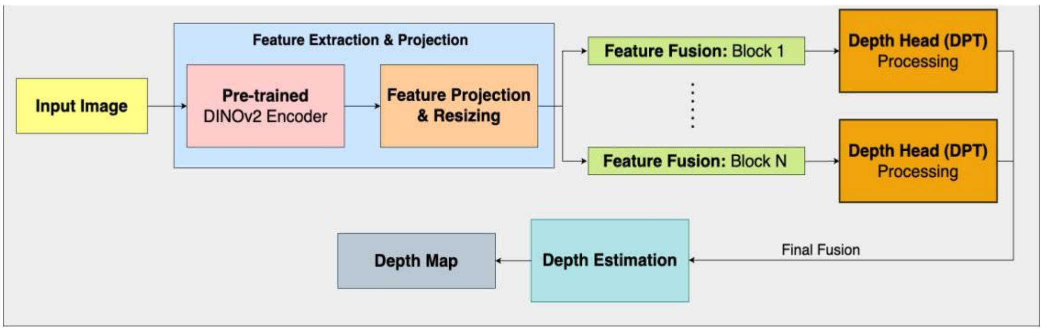

Input Image
Feature Extraction & Projection
Pre-trained DINOv2 Encoder
Feature Projection & Resizing
Feature Fusion: Block 1
Feature Fusion: Block N
Depth Head (DPT) Processing
Depth Head (DPT) Processing
Final Fusion
Depth Estimation
Depth Map

Fig.1. Architecture Of Proposed System

Monocular Depth Estimation

Monocular depth estimation involves predicting the depth of a scene from a single image. The system architecture typically consists of:

Encoder: A convolutional neural network (CNN) that extracts features from the input image.Thisnetwork typically consists of multiple layers, such as VGG, ResNet, or MobileNet, which are pre-trained on large datasets like ImageNet to capture general image features effectively.

Decoder: Another CNN that takes the features extracted by the encoder and generates a depth map. This network may consist of several up sampling and convolutional layers to progressively refine the depth information.

Skip Connections: To capture both low-level and high-level features, skip connections are often employed. These connections directly link corresponding layers from the encoder to the decoder, allowing the decoder to access fine-grained details from earlier stages of the encoding process.

Loss Function:Afunctionthatquantifiesthedifferencebetweenthepredicteddepthmapandthe ground truth depth map. Common loss functions used in monocular depth estimation include mean squared error (MSE) or structural similarity index (SSIM)

Training Data: A large dataset containing paired images and their corresponding depth maps is used to train the network. These datasets are often generated using depth sensors like LiDAR or stereo cameras, with additional data augmentation techniques applied to enhance the diversity of the training data.

By combining these components effectively, a monocular depth estimation system can accurately predict the depth of a scene from a single input image, enabling various applications such as augmented reality, autonomous driving, and robotics.

BACKGROUND SUBTRACTION

Background subtraction is a fundamental technique in image processing and computer vision used for detecting moving objects within a scene by separating them from the static background. Here's how it works:

Background Modeling: The process begins with creating a model of the background scene. This modelrepresentswhat thescenelookslikewithoutanyforegroundobjects. Itcan beconstructedusinga single frame (static background model) or a collection of frames (dynamic background model) captured over time. Techniques range from simple averaging of pixel values over time to more sophisticated methods such as Gaussian mixture models or deep learning-based approaches.

Foreground Detection: Once the background model is established, each new frame of the video or image sequence is compared with the background to identify pixels that differ significantly. These differing pixels are considered potential foreground pixels, indicating the presence of moving objects. The comparison can be done using simple intensity differences or more complex methods like color or texture analysis.

Foreground Segmentation: The detected foreground pixels are then segmented to delineate the boundaries of individual moving objects. This segmentation step helps isolate each object from its surroundings, making it easier to track and analyze.

Post-processing: To refine the results and remove noise or artifacts, post-processing techniques such as morphological operations (e.g., erosion, dilation) or connected component analysis may be applied. These operations help to smooth object boundaries, fill in gaps, and remove small isolated regions that are unlikely to be actual objects.

Object Tracking: Background subtraction is often used as a precursor to object tracking, where the detected foreground objects are followed over time to estimate their trajectories and predict their future positions. Tracking algorithms can employ techniques such as Kalman filtering, particle filtering, or deep learning-based methods to maintain object identity and handle occlusions or changes inappearance.

Background subtraction is commonly used in various applications such as surveillance, traffic monitoring, human-computer interaction, and video analysis. While it is effective in scenarios with relatively static backgrounds and well-defined foreground objects, it may encounter challenges in complex environments with dynamic backgrounds, changing lighting conditions, or occlusions. Advanced techniques, including adaptive background modeling and hybrid approaches combining background subtraction with other methods like optical flow or deep learning, are employed to address these challenges and improve detection accuracy and robustness.

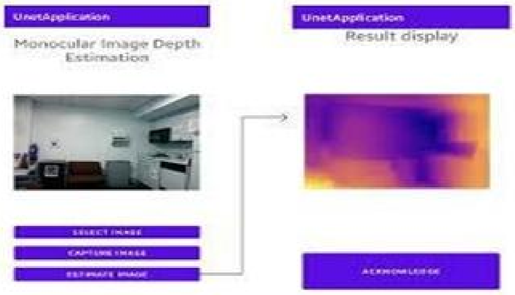


Fig.2. Background Subraction

MASKING

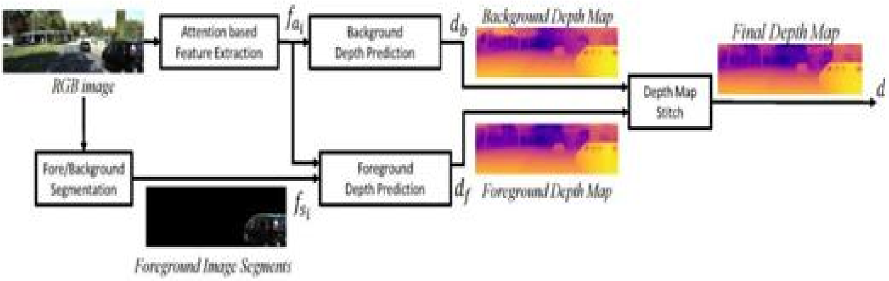

RGB image
Attention based Feature Extraction
f_{a_1}
Background Depth Prediction
d_b
Background Depth Map
Final Depth Map
Depth Map Stitch
d
Fore/Background Segmentation
Foreground Depth Prediction
d_f
Foreground Depth Map
f_{s_1}
Foreground Image Segments

Fig.3. Fore ground Masking

Maskinginimagedetectioninvolvesusingabinarymasktoisolateorextractspecificregionsofinterest within an image. Here's a brief explanation of how masking works:

1. **Definition of Mask**: A mask is essentially a binary image with the same dimensions as the original image, where each pixel is assigned a value of either 0 (black) or 1 (white). Pixels with a value of 1 in the mask represent the regions of interest (foreground), while pixels with a value of 0 represent the background.

2. **Application of Mask:** To apply the mask to the original image, a pixel-wise multiplication operation is performed between the mask and the original image. This operation results in a new image where pixels outside the masked region are set to zero (black), effectively "masking out" those areas and leaving only the regions specified by the mask intact.

3. **Isolation of Regions**: By using different masks, specific regions or objects within the image can be isolated or extracted. For example, a mask can be created to highlight only certain colors, shapes, or textures within the image,

effectively filtering out irrelevant information and focusing attention on the desired features.

4. **Combination of Masks**: Multiple masks can be combined using logical operations such as AND, OR, orXOR to createmorecomplex masks orto extract overlapping regions ofinterest. This allowsfor finer control over the selection and extraction of image elements based on various criteria.

5. **Usage in Object Detection**: In the context of object detection, masking can be used to segment objects from the background, making it easier to identify and analyze individual objects within the scene. Masking techniques are often employed in conjunction with other detection methods, such as edge detection, thresholding, or machine learning-based approaches, to refine object boundaries and improve detection accuracy.

Overall, masking is a powerful technique in image detection that enables selective processing and analysis of specific regions within an image, facilitating tasks such as object segmentation, feature extraction, and image enhancement.

IMAGE TRACKING

In monocular depth estimation, image tracking involves the process of following objects or features of interest across consecutive frames of a video sequence. Here's a brief explanation:

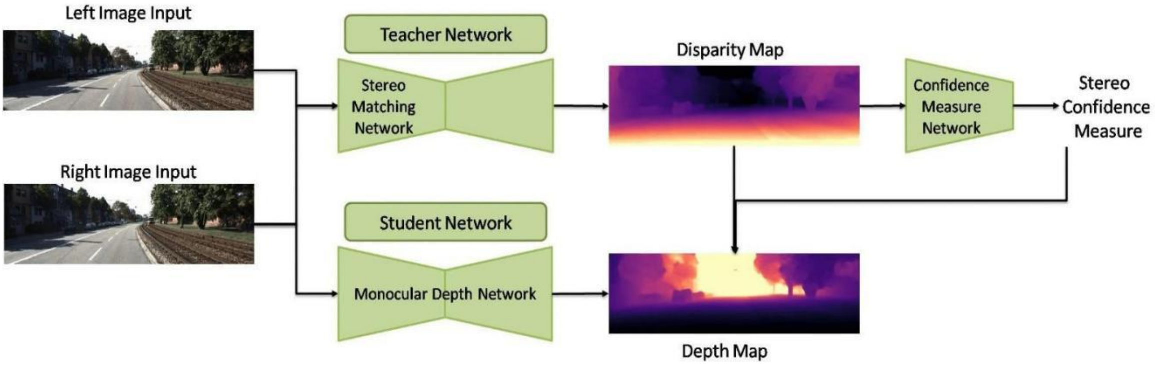


Fig:4 ImageTracking

InitialObjectDetection:Beforetrackingbegins,objectsorfeaturesofinterestare typically detected in the first frame of the video sequence using object detection or feature extraction techniques. This initial detection provides the starting point for tracking.

1. **Feature Matching**: In subsequent frames, the detected objects or features are matched with their corresponding counterparts from the previous frame. This matching process can be based on various similarity measures, such as spatial proximity, appearance similarity, or feature descriptors like SIFT (Scale-Invariant Feature Transform) or SURF (Speeded-Up Robust Features).

2. **Motion Estimation**: Once the corresponding features are identified, the motion between frames is estimatedtodeterminehow theobjectsor featureshavemoved.Thismotionestimationcanbeachieved

3. **Update and Refinement**: As new frames are processed, the object positions or feature locations are updatedbasedontheestimatedmotion.Additionally,trackingalgorithmsmayincorporatetechniquesto refine the tracking results, such as Kalman filtering, which predicts the object's future position based on its past trajectory and corrects any errors in the tracking process.

4. **Challenges and Considerations:** Image tracking in monocular depth estimation faces challenges such as occlusions, changes in object appearance, and camera motion. To address these challenges, tracking algorithms may employ strategies such as feature re-detection, adaptive model updating, or robust estimation techniques to maintain accurate tracking across varying conditions.

Overall, image tracking plays a crucial role in monocular depth estimation by enabling the consistent tracking of objects or features over time, which is essential for tasks such as scene reconstruction, motion analysis, and 3D mapping.

IMAGERECOGNITION

In monocular depth estimation, image recognition involves identifying and categorizing objects or scenes within a single image. Here's a brief explanation

Object Detection and Classification: Image recognition algorithms detect objects within an image andclassifytheminto predefinedcategories.Techniquessuchasconvolutionalneuralnetworks(CNNs) are commonly used for this task, where the network learns to recognize object features and classifythem based on learned patterns.

the algorithm for discovering novel object features builds upon previous iterations by incorporating advanced clustering, representation learning, and novelty detection techniques to enhance feature discovery efficiency, accuracy, and adaptability across various application domains.

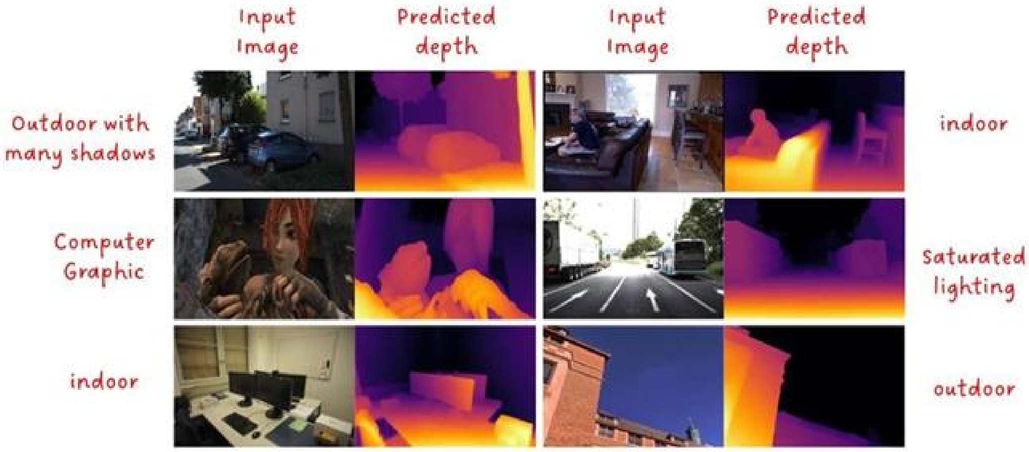
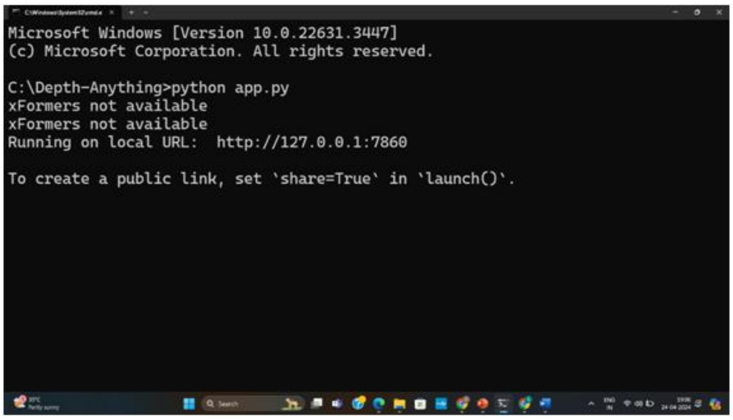


Fig.5. input image

RESULTS

The execution of the process will be explained clearly with the help of the continuous screenshots. The whole process in the execution is uploading a image after uploading image and we have to submit the image. After submitting, the system will automatically estimate the depth map through colors. This whole process is done in four simple steps. Each figure mentioned below are the simultaneous process of screening outputs.

LunchingPython app.py

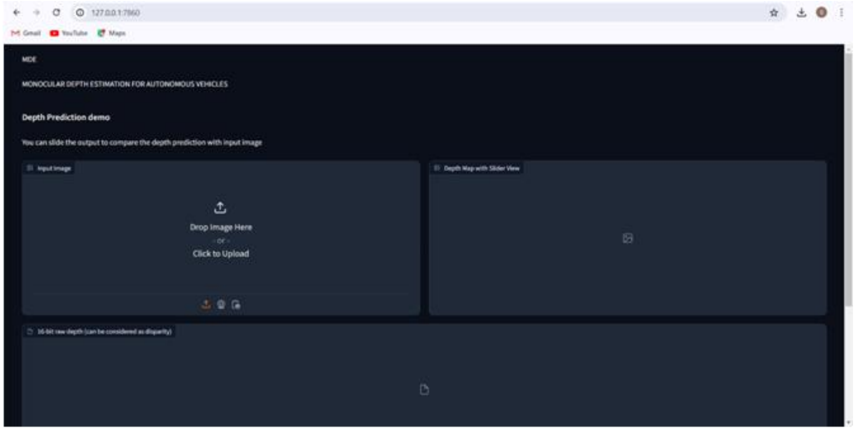


```
Microsoft Windows [Version 10.0.22631.3447]
(c) Microsoft Corporation. All rights reserved.

C:\Depth-Anything>python app.py
xFormers not available
xFormers not available
Running on local URL:  http://127.0.0.1:7860

To create a public link, set `share=True` in `launch()`.
```

opensaURLlink



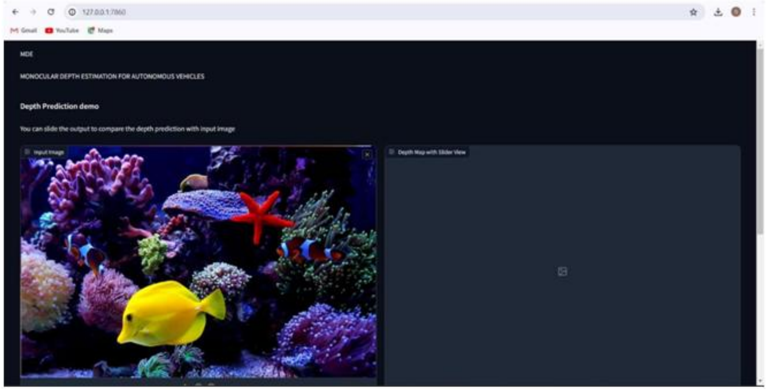
MDE

MONOCULAR DEPTH ESTIMATION FOR AUTONOMOUS VEHICLES

Depth Prediction demo

You can slide the output to compare the depth prediction with input image

Upload Image



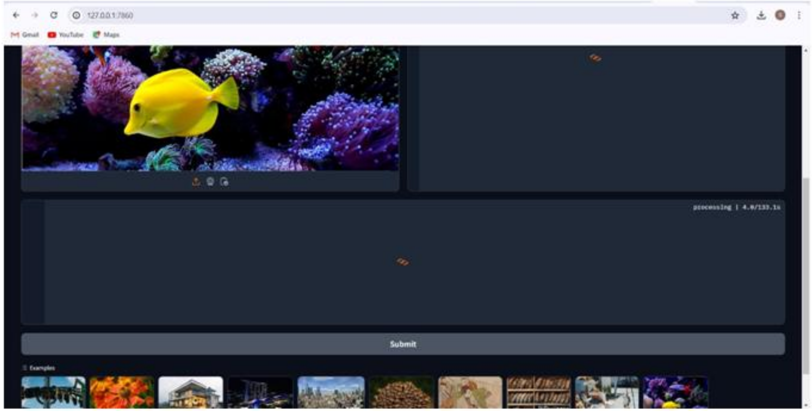
MDE

MONOCULAR DEPTH ESTIMATION FOR AUTONOMOUS VEHICLES

Depth Prediction demo

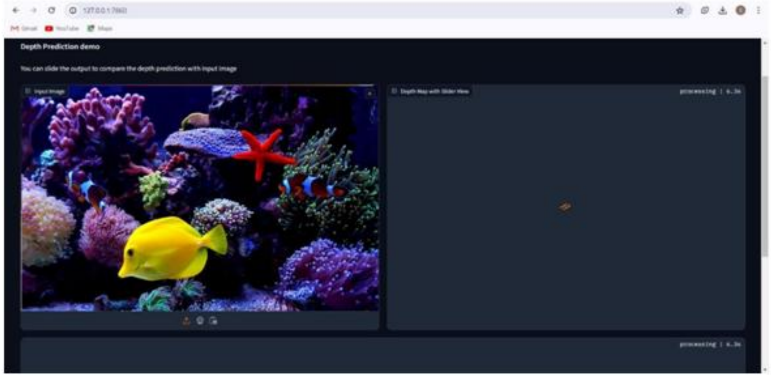
You can slide the output to compare the depth prediction with input image

Submit the Image



```
processing | 4.0/133.1s
```

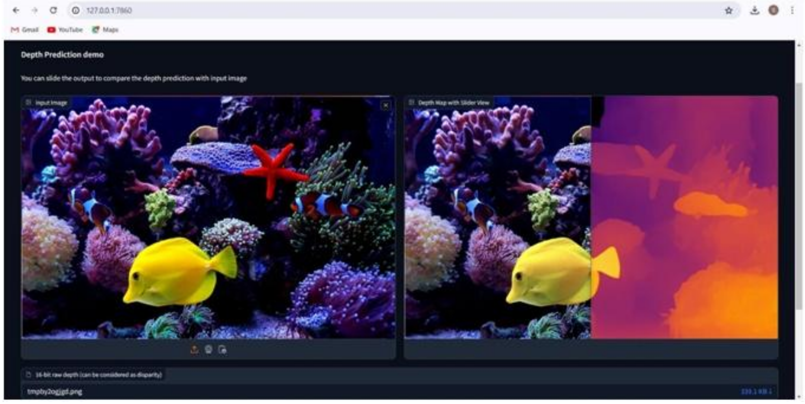
Image processing



Depth Prediction demo

You can slide the output to compare the depth prediction with input image

Depth Map with Slider View

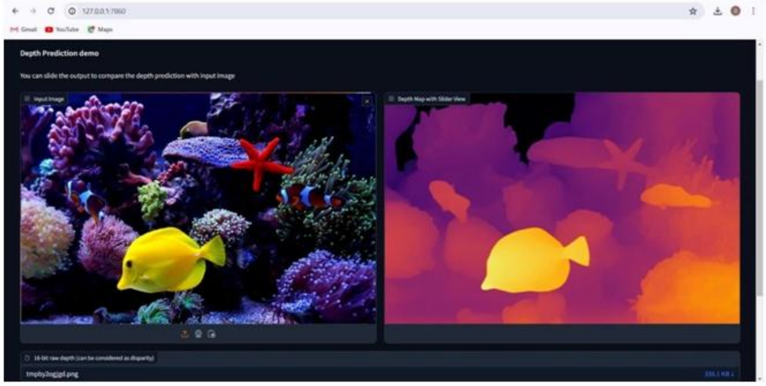


Depth Prediction demo

You can slide the output to compare the depth prediction with input image

127.0.0.1:7860
Gmail
YouTube
Maps
Input Image
Depth Map with Slider View
16-bit raw depth (can be considered as disparity)
tmpby2ogjgd.png

Final Depth Map



Depth Prediction demo

You can slide the output to compare the depth prediction with input image

127.0.0.1:7860
Gmail
YouTube
Maps
Input Image
Depth Map with Slider View
16-bit raw depth (can be considered as disparity)
tmpby2ogjgd.png

CONCLUSION

Depth anything presents a highly practical solution for robust monocular depth estimation, emphasizing the utilization of cheap and diverse unlabeled images.Throughtheimplementationoftwosimpleyetpowerfulstrategies,namelyposing amorechallengingoptimizationtargetduringlearningandpreservingrichsemanticpriors from pre-trained models, depth anything achieves remarkable zero-shot depth estimation performance. Moreover, it serves as a promising initialization for downstream tasks such as metric depth estimation and semantic segmentation. By leveraging these strategies, depth anything not only demonstrates impressive generalization ability but also sets new benchmarks in metric depth estimation. This work underscores the importance of innovative approaches to dataset scaling-up and highlights the potential of unlabeled data in advancing the field of monocular depth estimation.

FUTURE ENHANCEMENTS

In future work, enhancing Depth Anything could involve exploring novel techniques for incorporating temporal information to handle dynamic scenes more effectively, thereby improving its performance in scenarios with moving objects or changing environments. Additionally, investigating methods to leverage self-supervised learning approaches or unsupervised domain adaptation could further enhance the model's ability to generalize across diverse datasets and domains, ultimately advancing its practicality and robustness for real-world applications in robotics, autonomous driving, and virtual reality.

References

1. Manuel L´opezAntequera, Pau Gargallo, Markus Hofinger, Samuel RotaBul`o, Yubin Kuang, and Peter Kontschieder. Mapillary planet-scale depth dataset. In ECCV, 2020. 7, 9
2. HangboBao, Li Dong, SonghaoPiao, and Furu Wei. Beit: Bert pre-training of image transformers. In ICLR, 2022. 7
3. Shariq Farooq Bhat, IbraheemAlhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In CVPR, 2021. 2, 6
4. Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias M¨uller. Zoedepth: Zero – shot transfer by combining relative and metric depth. arXiv:2302.12288, 2023. 2, 6, 7, 9
5. Reiner Birkl, Diana Wofk, and Matthias M¨uller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. arXiv:2307.14460, 2023. 2, 3, 5, 7, 8, 9, 10, 11
6. Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. arXiv:2108.07258, 2021. 1
7. Daniel J Butler, JonasWulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In ECCV, 2012. 5, 7, 9
8. Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. arXiv:2001.10773, 2020. 7
9. Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu- Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In CVPR, 2019. 2, 4
10. Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Singleimage depth perception in the wild. In NeurIPS, 2016