# A Self-Supervised Monocular Depth Estimation Approach Based on UAV Aerial Images

**Conference Paper** · September 2022

DOI: 10.1109/DASC55683.2022.9925733

Yuhang Zhang
Nanyang Technological University

Chen Lv
Nanyang Technological University

# A self-supervised monocular depth estimation approach based on UAV aerial images

Zhang, Yuhang; Yu, Qing; Low Kin Huat; Lv, Chen

2022

Zhang, Y., Yu, Q., Low Kin Huat & Lv, C. (2022). A self-supervised monocular depth estimation approach based on UAV aerial images. 2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC). https://dx.doi.org/10.1109/DASC55683.2022.9925733

https://hdl.handle.net/10356/162468

https://doi.org/10.1109/DASC55683.2022.9925733

# A Self-Supervised Monocular Depth Estimation Approach Based on UAV Aerial Images

Yuhang Zhang, Qing Yu, Kin Huat Low*, Chen Lv
School of Mechanical and Aerospace Engineering
Nanyang Technological University, Singapore
yuhang002@e.ntu.edu.sg, qing001@e.ntu.edu.sg, mkhlow@ntu.edu.sg*, lyuchen@ntu.edu.sg

***Abstract*—The Unmanned Aerial Vehicles (UAVs) have gained increasing attention recently, and depth estimation is one of the essential tasks for the safe operation of UAVs, especially for drones at low altitudes. Considering the limitations of UAVs' size and payload, innovative methods combined with deep learning techniques have taken the place of traditional sensors to become the mainstream for predicting per-pixel depth information.**

**Since supervised depth estimation methods require a massive amount of depth ground truth as the supervisory signal. This article proposes an unsupervised framework to tackle the issue of predicting the depth map given a sequence of monocular images. Our model can solve the problem of scale ambiguity by training the depth subnetwork jointly with the pose subnetwork. Moreover, we introduce a modified loss function that utilizes a weighted photometric loss combined with the edge-aware smoothness loss to optimize the training. The evaluation results are compared with the model without weighted loss and other unsupervised monocular depth estimation models (Monodepth and Monodepth2). Our model shows better performance than the others, indicating potential assistance in enhancing the capability of UAVs to estimate distance with the surrounding environment.**

***Index Terms*—UAV, self-supervised learning, monocular depth estimation, aerial images, multi-scale upsampling**

## I. Introduction

In recent years, due to the factors of flexibility and versatility, Unmanned Aerial Vehicles (UAVs) are developing rapidly in a wide range of areas, such as rescue, surveillance, and navigation. Since they can be flown in treacherous terrains and inaccessible areas together with real-time data acquisition and fast processing [1], UAVs have emerged as powerful platforms for the collection of high-resolution aerial images. To ensure a safe flight, depth estimation is indispensable for drone operations.

Depth Estimation refers to the process of predicting the metric depth value of each pixel from the given input image(s). Depth information is indispensable for 3D reconstruction tasks, which play a critical role in operating UAVs remotely. Multiple methods have been proposed to tackle this issue, such as LIDAR [2], RGB-D cameras [3] and Radar [3]. Unfortunately, for some small-sized UAVs [4], LIDAR and Radar suffer from the exorbitant price and unaffordable weight. Oppositely, RGB cameras are relatively inexpensive and lightweight. More noteworthy is that they can offer higher measurement densities, making them more qualified for UAVs applications with limited total payload.

In previous years, several attempts [5–8] have been made to predict depth from stereo images. Compared to monocular sequences, stereo methods are more accurate. Nevertheless, they need more resources to calibrate when acquiring data. Besides, the performance of stereo-based methods is directly related to the baseline distance between lenses [9]. Specifically speaking, when applying stereo-based methods to small-sized UAVs, estimation accuracy will be undermined by the short baseline distance limited by the weight and size of UAVs.

Therefore, for drone applications, it is more desirable to obtain depth information from a single RGB image, which is also defined as Monocular Depth Estimation (MDE). Humans perform well in this task by seamlessly combining cues like perspective, object sizes, and occlusions. However, MDE is an inherently ambiguous problem because a numerous number of various 3D scenes can be projected into the identical 2D coordinates. To cope with this, previous approaches are dedicated to searching for meaningful monocular cues, such as texture variations, color/haze, gradients, etc. Currently, Deep Learning (DL) has dominated many areas of computer vision, such as object detection [10] and semantic segmentation [11]. MDE has gained more attraction with the success of various deep learning-based models. Some current researches [12–15] reveal that depth information can be learned by using convolutional neural networks (CNNs) to decrease the errors between the predicted results and the labeled ground truth. Although these methods have achieved great success, they are only available for applications where sufficient image datasets and their corresponding depth ground truth are accessible.

In this paper, considering the challenge of acquiring a large amount of depth ground truth, a self-supervised method is proposed to predict depth information directly from monocular sequences. The methodologies of this paper are threefold:

- This self-supervised framework comprises two parts: the depth estimation and the pose estimation subnetworks. The values generated by these two subnetworks are utilized jointly in the forward process to infer the depth map given the image sequences captured by a monocular camera.

* Corresponding author

- Using the output from the networks to reconstruct the projection relationships of pixels between sequential images and the photometric loss is applied as the supervision signal to train the networks.
- Regarding the scale ambiguity problem that commonly exists in self-supervised learning, a modified loss function that combines photometric loss with edge-aware smoothness loss is proposed to tackle this issue. Since the weights of depth and pose networks are jointly optimized during the training, the scale consistency of the depth network is improved by this pose constraint at the same time.

The model is trained on the low altitude UAVs dataset [16] , and experimental results are compared with other unsupervised models, showing a potential to make contributions to the safe operations of UAVs at low altitudes.

## II. Related Works

After emerging as the platforms for defense and military purposes initially, small-scaled UAVs such as quadcopters are still being deployed in other innovative fields. Examples include, but are not limited to: forest fire detection [17], traffic surveillance [18], air pollution monitoring [19], and wildlife protection [20]. To better complete high-demanding tasks in complex environments that are dangerous to human beings, autonomous maneuvers like hovering and gliding are strictly required in drone operations, especially for UAVs at low altitudes. Therefore, perceiving and building a comprehensive 3D representation of the surrounding environment is the priority for UAVs' autonomous flying systems. An ocean of methods has been proposed to obtain depth information which is treated as an important part of environment perception and reconstruction.

### *A. Motion-Based Depth Estimation*

Motion-based depth estimation, such as Structure from Motion (SfM), takes images from different angles as inputs to cope with camera motion recovery and 3D environment modelling [21]. Given a set of images captured from different perspectives, the pipeline of SfM always starts with feature extraction and matching, followed by geometric verification. After removing the incorrectly matched outliers, left-matched features are tracked to calculate the 3D point cloud map, which can be transformed into the corresponding depth map.Ping Li [22] uses a factorization-based SfM approach to create the depth map from uncalibrated video sequences of static scenes. The proposed approach divides the motion and structure recovery part into two sub-steps, including the projective and the Euclidian reconstruction, to improve the depth map quality in the textureless environment. Li Ding [23] presents a framework combining SfM with lidar scan to estimate dense depth maps precisely. This framework increases the efficiency of the traditional SfM method by tracking feature correspondences following a 3D-to-2D alignment, and the results perform well on high-resolution imagery.

Feature matching is a decisive factor in improving the accuracy of depth values generated by SfM methods. Furthermore, the phenomenon of insufficient features is prone to appear when encountering environments with less texture or low contrast. Consequently, available SfM methods can only generate sparse depth maps in most cases. These sparse maps are not qualified for drone applications in which a dense depth map is required for UAVs to complete complex maneuvers.

### *B. Learning-Based Depth Estimation*

With the continuous growth of deep learning, deep learning-based methods have displayed their surpassing performance in image processing. Over the last several years, there have been some impressive breakthroughs in learning-based monocular depth estimation models. Learning-based depth estimation approaches can be categorized into two major parts from the availability of ground truth: supervised and unsupervised.

*1) Supervised Depth Estimation*

The pipeline of supervised depth estimation approaches can be explained as below. Supervised learning aims to penalize the errors between estimated and true depth values using deep neural networks. The MDE network integrates an input image I and the matching ground truth depth $D^*$, and utilizes a loss function $\mathcal{L}_2\left(D, D^*\right)$ as the supervisory indicator to learn the depth information of scenes. The network converges when $D$ is close to $D^*$ to the greatest extent.

$$\mathcal{L}_2\left(D, D^*\right)=\frac{1}{N} \sum_i^N\left\|D-D^*\right\|_2^2 \tag{1}$$

where $D$ is the predicted depth map.

Eigen [12] first employed a supervised learning algorithm to solve the MDE problem to the best of our knowledge. In this algorithm, two CNNs (the overall rough network and the local fine-grained network) are employed to address the uncertainty and ambiguity coming from the overall scale. In addition, the difference generated by various scale is introduced as the loss function to help minimize the error between the prediction and ground truth, which proves its remarkable performance of MDE on the KITTI [24] and NYU [25] datasets. Eigen [13] designs a single multiscale convolutional network architecture capable of adapting to multiple tasks such as depth prediction and semantic labeling with only tiny modifications. This network works by reducing the gradient error between the predicted and reference values in the horizontal and vertical directions. Experimental results show that this model can generate real-time outputs (∼30Hz).

Such an approach works well on images with short-range depth values and richly textured backgrounds. Supervised-based solutions for aerial images, on the other hand, are relatively troublesome due to the lack of fixed structures (degress of freedom for the drone are much more than self-driving vehicles). To bridge the gap, Miclea and Nedevschi [26] bring out a CNN incorporating a novel scene digging module and an innovative softmax transformation layer to achieve a greater convergence in aerial scenarios. This architecture introduces an optimal feature extractor with a creative upsampling subnetwork and combines both ordinal regression (that better accounts for areas with less texture variation) and classification (that performs better on stand-alone objects) as loss function to train the model. The model is trained using images from the MidAir [27] dataset and validated on synthetically generated and authentic aerial images. Although the model shows promising performance in the validation scenarios, it is still not general enough to other real-life aerial scenarios.

### 2) *Unsupervised Depth Estimation*

Considering the possibility of acquiring a sufficient amount of ground truth, which is also expensive, one significant improvement in recent years was introducing unsupervised models for the MDE problem. The training process of unsupervised models is solely governed by the information available in the RGB images. The pipeline of unsupervised depth estimation approaches can be depicted as below. Unsupervised methods take the geometric restrictions of objects in adjacent frames as the supervision indicator. More specifically, unsupervised models always take a monocular video sequence or a stereo pair as inputs (in which one image is the target image and another one is the reference image), along with camera intrinsics and the spatial transformations between frames. The focus of the training goal is straightforward: after getting the depth map predicted by the network, points generated based on the inferred depth in the target image can be warped to the reference image to get the reconstructed image combined with the camera movement information (see Equation 2). Due to the inaccurate depth prediction results, there is photometric error between the original and reconstructed images, the network can be trained based on the photometric error instead of the ground truth.

$$P_n \sim \mathbf{K} T_{n-1 \rightarrow n} D_{n-1}\left(P_{n-1}\right) \mathbf{K}^{-1} P_{n-1} \tag{2}$$

where $P_n$ represents the pixel on current frame $I_n$, and $P_{n-1}$ denotes the matching pixel of $P_n$ on the previous frame $I_{n-1}$. $\mathbf{K}$ is the internal parameter of the camera, which is assumed to be a known constant. $T_{n-1 \rightarrow n}$ stands for the self movement matrix between $I_{n-1}$ and $I_n$, and $D_{n-1}\left(P_{n-1}\right)$ refers to the depth value at pixel $P_{n-1}$.

To the best of our knowledge, Garg [28] introduces the first unsupervised framework for the MDE problem. In this framework, stereo pairs with known camera motion are employed to exploit the non-linear functional relationship between image and depth map. There is a color deviation between the original image and the reconstructed image, which is employed to monitor the process of updating the weights of the network. Godard [29] presents a novel CNN network to improve the quality of depth images. This work uses calibrated stereo pairs that include left and right color images for training. Both disparities (left-to-right and right-to-left) can be inferred simultaneously from the left input image. Moreover, error in disparity consistency of a stereo pair is proposed as the loss function to enforce mutual consistency between the left-view disparity and the projected right-view disparity map. Although this method is not suitable for single-view datasets, it still outperforms many other supervised methods. Godard [30] tries to robustly handle occlusions and reduce visual artifacts in the unsupervised methods. In his method, a minimum reprojection loss and an auto-masking loss are proposed to filter out occluded pixels and objects moving at the identical velocity as the camera. Moreover, a multiscale estimation method is introduced to contribute to the convergence speed of the network. This method performs well on both monocular videos and stereo pairs.

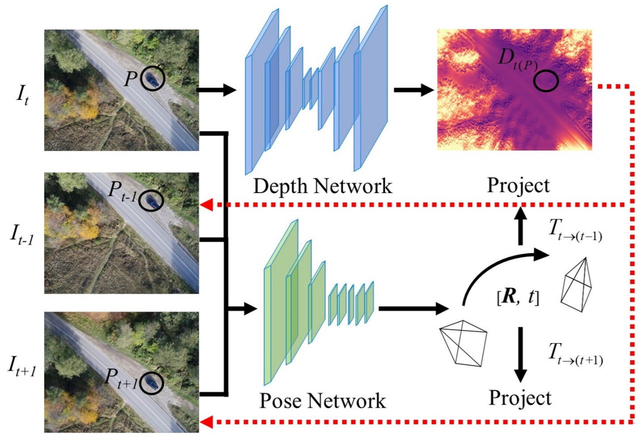


Fig. 1. Overview of the pipeline of our self-supervised depth estimation model

Although a large proportion of approaches handle the MDE problem in the ground-related scenarios (for automotive and indoor), learning-based depth estimation for aerial environments is still an unexplored field, with many new algorithms to explore. Several researches [31][32] are attempting to create datasets by using a simulator to generate synthetic images, which are categorized into two parts (low altitude range and high altitude range). While such datasets can provide a reference outdoor environment, there is an overall mismatch with the natural atmospheric environment (from the aspects of illumination and environment appearance). Such a research gap highlights the challenges of the aerial scenario, most originating from the unconstructed ego-motion and complex backgrounding texture.

## III. Methodology

### A. *Overview*

This section proposes a model for jointly inferring a depth map and the camera motion information between the continuous frames from an unlabeled video sequence. The model comprises two parts: a subnetwork designed for depth prediction and a subnetwork designed for pose estimation. Fig. 1 depicts the overall framework of our model. We denote the input frame of the depth estimation subnetwork as the target image $I_t$, and images at the previous moment or later as reference images $I_{t-1}$ or $I_{t+1}$. The depth estimation subnetwork aims to to generate the depth map $D_t$ from $I_t$. The pose

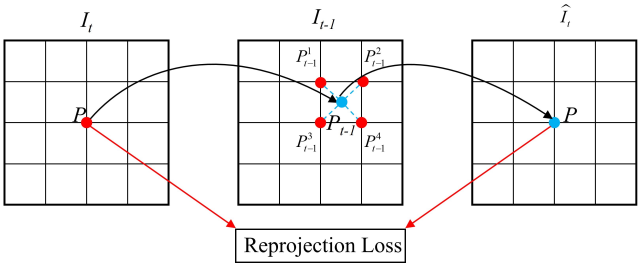


Fig. 2. An illustration of image warping process. After projecting the pixel point in $I_t$ to $I_{t-1}$, bilinear sampler is employed to calculate the 2d coordinates of the projected pixel point in the $I_{t-1}$ to reconstruct the image. The differences between target and reconstructed images are considered as reprojection loss.

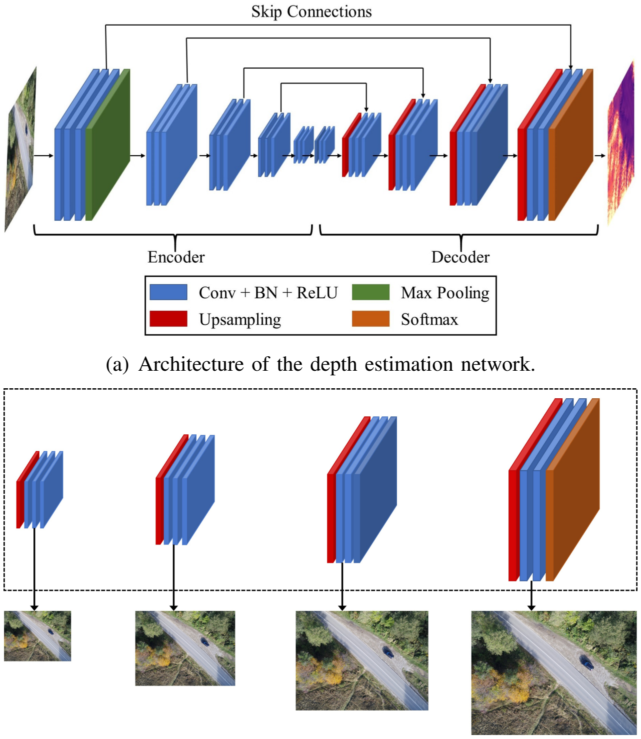
estimation subnetwork aims to output the self-motion matrix $T_{t\to(t-1)}$ or $T_{t\to(t+1)}$ by taking a pair of target and reference images ($I_{t-1}$, $I_t$), or ($I_t$, $I_{t+1}$) as input. Subsequently, the reference images ($I_{t-1}$ or $I_{t+1}$) can be warped into the target image ($I_t$) to reconstruct the target view (see Fig. 2). By utilizing the photometric loss between the target image and reconstructed image as supervision, we are capable of training the entire model in a self-supervised manner.

### B. Depth Estimation Subnetwork

Since U-Net [33] has the capability to localize each pixel on the image and distinguish specific patterns precisely, it is commonly utilized for pixel-level prediction of depth estimation. Therefore, we employ a modified version of this typical structure to create the depth subnetwork. The architecture of the depth subnetwork is shown in Fig. 3(a). Similar to the traditional U-net, our network consists of an encoder, a decoder, and four skip connections. The encoder comprises convolution blocks, batch normalization layers, and a Max-pooling layer, all of which extract different sizes of depth features from input images. Unlike the traditional U-Net, our convolution blocks are built based on ResNet50 [34], and the output channels of each convolution block are 32, 64, 128, 256, and 512, respectively, removing the convolution block related to 1024 output channels to accelerate depth estimation. The decoder comprises the same convolution blocks as the encoder, upsampling layers, and a Softmax layer to restore the extracted features to the original size. The output channels of the first four convolution blocks in the decoder are 512, 256, 128, and 64, respectively, and the output channels in the last convolution blocks are 32 and 1. The function of skip connections is to combine the information in the shallower and the deeper layers, making the network use fine-grained features learned in the encoder to predict depth in the decoder.

### C. Multi-Scale Upsampling

To comprehensively combine the features extracted from all decoder layers, which are conducive to predicting the depth map more accurately, we propose a multi-scale upsampling estimation method to calculate the weighted loss sum in all decoder layers. More specifically, it is shown in Fig. 3(b) that images with four different resolutions are produced during the decoder. Losses obtained at different scale in the decoder can be calculated compared with their corresponding reconstructed images respectively. Therefore, the total loss is the combination of each photometric error at the resolution of each decoder layer.



(a) Architecture of the depth estimation network.

(b) Reconstructed images at different resolutions in the decoder.

Fig. 3. Framework of the depth estimation network.

### D. Pose Estimation Subnetwork

As for the pose estimation subnetwork, we also use a modified ResNet50 as the encoder. The encoder accepts two adjacent frames (six channels) as input to extract features, and the output channels of the convolution block in the encoder increase from 16 to 512 exponentially. The decoder comprises five convolution blocks followed by a global average pooling layer, and the final output channels of the decoder are 6, corresponding to the 6-DoF relative camera pose. The architecture of the pose estimation subnetworks is shown in Fig. 4.

### E. Automatic Mask

The network proposed in this article must meet the prerequisites of a moving camera and a stationary background. However, these assumptions are difficult to meet in real-world application scenarios. For example, when UAVs are required to perform hovering operations in textureless environments or there is a moving object in the current scene, prediction can degenerate significantly. To solve this problem, we employ an automatic mask $\mu$ to exclude the loss of pixels where the reprojection error of the reconstructed image $I_{r\to t}$ is higher than that of the unwarped reference image $I_r$. Since a static

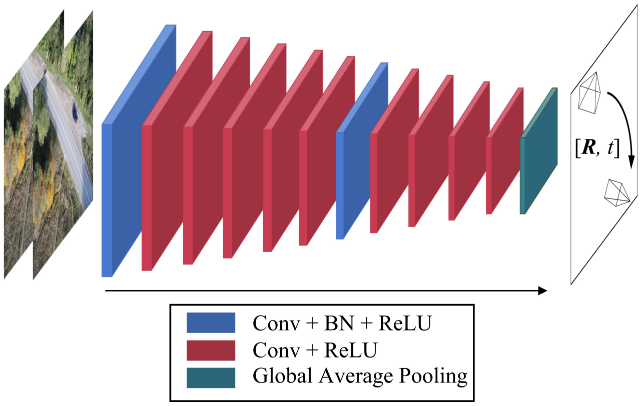


Fig. 4. Architecture of the pose estimation network

camera or an object moving at a similar velocity as the camera always produces identical pixels between adjoining frames, this formulation can filter out these pixels that break down the previous assumptions.

$$\mu = \left[\min_r pe\left(I_t, I_{r\to t}\right) < \min_r pe\left(I_t, I_r\right)\right] \tag{3}$$

where [ ] is the Iverson bracket, then $\mu$ is 1 when the conditions are determined to be ture.

### F. Loss Function

Our model is guided to update the network parameters in the direction of minimizing photometric reprojection error. We denote the relative pose between the target image $I_t$, and the reference image $I_r$, ($I_{t-1}$ or $I_{t+1}$,) as $T_{t\to r}$. Our model tends to establish a dense depth map $D_t$ that generates the minimum photometric reprojection error $L_p$. Initially, the reconstructed image $I_{r\to t}$ can be obtained utilizing the following formulation:

$$I_{r\to t} = I_r \left\langle \operatorname{proj}\left(D_t, T_{t\to r}, K\right)\right\rangle \tag{4}$$

where $\operatorname{proj}()$ are the resulting 2D coordinates of the projected depth $D_t$ in $I_r$, $\langle\ \rangle$ is the bilinear sampling operator, and $K$ is the camera intrinsic matrix.

Minimum photometric reprojection loss $L_p$ is adopted as one of the loss indicators between $I_t$ and $I_{t\to r}$, the formulation can be expressed as:

$$L_p = \min_r pe\left(I_t, I_{r\to t}\right) \tag{5}$$

where $pe$ represents the photometric reprojection loss and $\min_r$ denotes the minimum value of $pe$.

Following the methods in [30], we build $pe$ based on L1 distance and SSIM:

$$pe\left(I_{r\to t}, I_t\right) = \frac{\alpha}{2}\left(1 - \operatorname{SSIM}\left(I_{r\to t}, I_t\right)\right) + (1-\alpha)\left\|I_{r\to t} - I_t\right\|_1 \tag{6}$$

where $\alpha$ is 0.85, $\left\|I_{r\to t} - I_t\right\|_1$ represents the $\mathrm{L}_1$-norm, and SSIM refers to the structural similarity between $I_{r\to t}$ and $I_t$. SSIM can be calculated utilizing the following formulation:

$$\operatorname{SSIM}\left(I_t, I_{r\to t}\right) = \frac{\left(2u_{I_t}u_{I_{r\to t}} + c_1\right)\left(2\sigma_{I_t I_{r\to t}} + c_2\right)}{\left(u_{I_t}^2 + u_{I_{r\to t}}^2 + c_1\right)\left(\sigma_{I_t}^2 + \sigma_{I_{r\to t}}^2 + c_2\right)} \tag{7}$$

where $c_1$ and $c_2$ are small numbers to protect the denominator from being 0, $u_{I_t}$, $u_{I_{r\to t}}$, $\sigma_{I_t}^2$, $\sigma_{I_{r\to t}}^2$ and $\sigma_{I_t I_{r\to t}}$ denotes the mean values, variance, and covariance of all pixels involved in the target and reference images respectively.

Subsequently, the edge-aware smoothness loss $L_s$ is introduced to form the overall loss function together with the photometric reprojection loss:

$$L_s = \left|\partial_x d_t^*\right| e^{-\left|\partial_x I_t\right|} + \left|\partial_y d_t^*\right|^{-\left|\partial_y I_t\right|} \tag{8}$$

where $d_t^* = d_t/\bar{d}_t$ is the inverse depth normalized by the mean value.

Since the features extracted in the low-resolution layers of the decoder reveal less about the contour of the depth map, which may lead to inaccurate network prediction, assigning the same weights to the low-resolution and high-resolution layers tends to result in the divergence of the network. Therefore, the overall loss function can be written as follows:

$$L = \sum_i w_i\left(\mu L_{p,i} + \lambda L_{s,i}\right) \tag{9}$$

where subscript $i$ represents the layers of the decoder at different resolutions, $w_i$ denotes the weights determined by the resolution, and $\lambda$ is the smoothing coefficient.

## IV. Experimental Results

### A. Dataset

WildUAV [16] is employed as the training dataset. WildUAV raw dataset contains over 1500 high-resolution (5280 × 3956) images and their corresponding depth ground truth data is provided jointly with camera intrinsics. The scenarios in the captured images contain both unstructured forest areas with various landforms and structured objects such as driveways, vehicles and pedestrians. Some samples of the dataset are shown in Fig. 5.

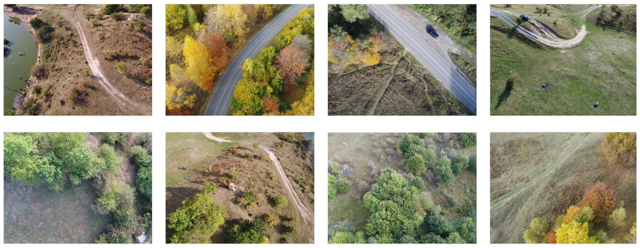


Fig. 5. Samples of the WildUAV dataset [16]

Although WildUAV comprises high-resolution images with rich information, the total number of images is insufficient to avoid overfitting. Therefore, we pre-processed the images before training. We crop ten sub-images from the center part

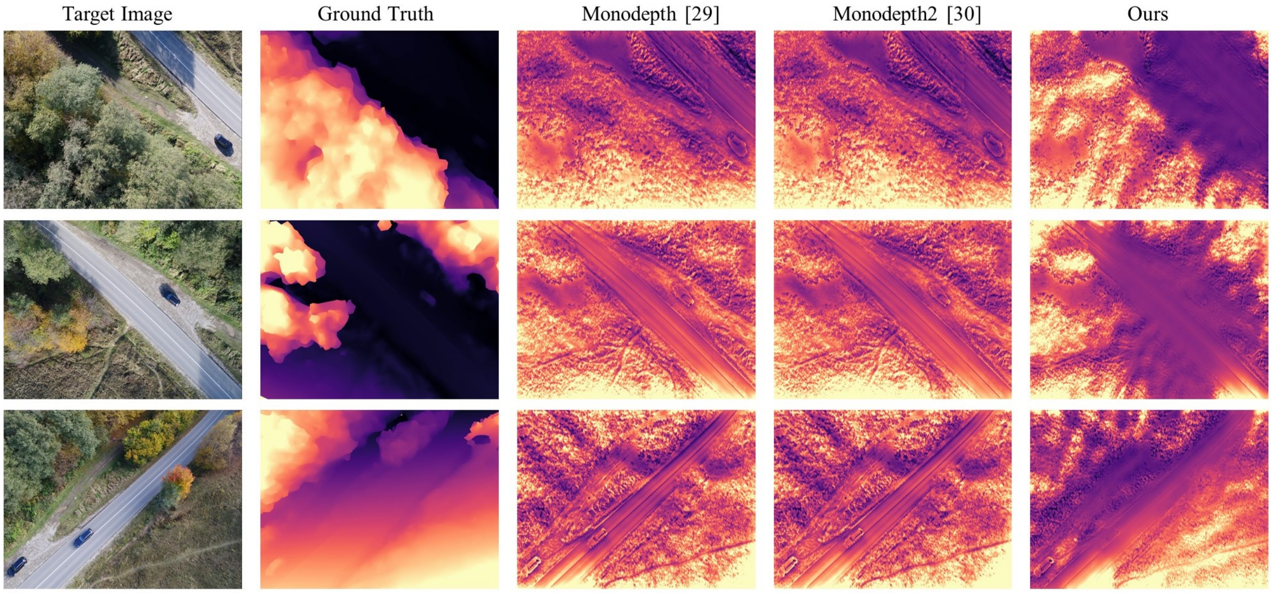


Fig. 6. Estimation results on the WildUAV dataset. Compared with Monodepth and Monodepth2, our model produces a more precise contour of the depth map, especially the part of the driveways. Since the two monocular depth estimation models used for comparison apply shallower networks than ours, they are prone to fail when encountering environments with less texture and longer distance in UAV application scenarios. Combined deeper networks with the weighted loss, our model can better adapt to textureless environments and more accurately predict depth information for densely wooded areas and sparse driveways.

TABLE I
THE EVALUATION RESULTS OF SEVERAL MONOCULAR DEPTH ESTIMATION MODELS ON WILDUAV

| Method | Loss | | | Errors | | | | Accuracy (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Photometric | Smoothness | Weighted | Abs Rel | Sq Rel | RMSE | RMSE log | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
| Monodepth [29] | ✓ | ✓ | ✗ | 0.187 | 1.503 | 6.133 | 0.250 | 0.795 | 0.928 | 0.969 |
| Monodepth2 [30] | ✓ | ✓ | ✗ | 0.174 | 1.435 | 5.786 | 0.221 | 0.831 | 0.939 | 0.974 |
| Ours | ✓ | ✓ | ✗ | 0.182 | 1.449 | 5.775 | 0.232 | 0.828 | 0.934 | 0.972 |
| Ours | ✓ | ✓ | ✓ | **0.149** | **1.219** | **5.246** | **0.216** | **0.857** | **0.941** | **0.978** |

for each image because it typically contains richer texture information than the boundary part. Subsequently, every sub-image is linked with its adjacent frame to get ten sequences. Then data segmentation is performed on these sequences to obtain the final training and validation sets. Eventually, the training set consists of 11961 images and the validation set consists of 1329 images.

### *B. Implementation Details*

We utilize only monocular image sequences as the training set and the network is trained in the development environment of Ubuntu 20.04, Pytorch 1.8.2 and CUDA 11.3. The details of the parameter setting are as follows: we set $\lambda$ = 0.001, $w$ = $\{$0.125, 0.25, 0.5, 1$\}$ accordingly. Our network is trained for 30 epochs with a batch size of 8. We utilize a learning rate of $2 \times 10^{-4}$ for the first 25 epochs and $10^{-4}$ for the rest. It takes about 6 hours to finish the monocular training on a computer equipped with four NVIDIA T4 Tensor Core GPUs (each with 2560 CUDA cores and 16 GB of memory) and NVIDIA T4 Tensor Core GPU.

### *C. Performance Evaluation Metrics*

To quantitatively evaluate the prediction results,several commonly used evaluation metrics in prior works, including absolute relative difference (Abs Rel), square relative difference (Sq Rel), the linear root mean square error (RMSE), the logarithmic root mean square error (LG RMSE), and accuracy under a threshold, are employed as indicators of judging the performance of models. Their corresponding formulas can be written as follows:

$$\text{Abs Rel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i - D_i^*|}{D_i^*} \tag{10}$$

$$\text{Sq Rel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i - D_i^*|^2}{D_i^*} \tag{11}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|D_i - D_i^*|^2} \tag{12}$$

$$\text{LG RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\lg D_i - \lg D_i{}^*|^2} \tag{13}$$

$$\text{Accuracy} : \max\left(\frac{D_i}{D_i^*}, \frac{D_i^*}{D_i}\right) = \delta < T \tag{14}$$

In the above formulas, $N$ is the total number of pixels in the target image, $D_i$ and $D_i^*$ denote the estimated depth and the depth ground truth, and $T$ is the artificially set thresholds ($1.25$, $1.25^2$, $1.25^3$).

### D. Results Analysis

We compare the estimation results with Monodepth [29] and Monodepth2 [30], both of which are unsupervised monocular depth estimation models showing great results on the KITTI [24] dataset. The comparison of the estimation results are shown in Fig. 6. It can be seen from Fig. 6 that our model produces a sharper depth map than the comparative models. Compared with the depth ground truth, our model performs better on the depth estimation of the drive lanes, while Monodepth and Monodepth2 show weaker performances in predicting their depth. We speculate that the introduction of deeper networks and weighted loss brings more promising prediction results. To further illustrate the model's performance, we also evaluated the model without introducing weight loss and Table I reveals the numerical results of the evaluation. In this table, better performance is characterized by smaller errors and higher accuracy, and the results show that weighted loss does have a positive impact on the model's training process. We infer that after weighting the loss of the shallower layers in the decoder, which provide less contour information of the depth map, the total loss function is less contaminated then the convergence of the model becomes better.

## V. Conclusion

In this article, a self-supervised depth estimation model is proposed to predict the depth map from images captured by UAVs at low altitudes. This model comprises a depth and a pose estimation subnetworks which are trained jointly to predict depth given a sequence of monocular images. This model utilized a modified loss function that weights the different layers' photometric loss according to their corresponding resolutions. Subsequently, this model is trained on a low altitude UAV dataset, and evaluation results are compared with other unsupervised monocular depth estimation models (Monodepth and Monodepth2) to manifest its superiority. The results show that the model tends to produce better results with the introduction of weighted reconstruction loss. This self-supervised monocular depth estimation model can make a contribution to the safe UAV operations.

## VI. Future Work

The monocular depth estimation algorithm proposed in this paper is applied to textureless wilderness scenes. Furthermore, safety and hazard analysis for UAVs should also be extended to the complex urban environments, which usually refers to the issue of Unmanned Aircraft System Traffic Management (UTM) [35][36]. Compared to wild environments, urban-like environments are typically rich in texture and occlusion, bringing new challenges to the monocular depth estimation of UAVs. We will focus on the future problem of UAV depth estimation in complex urban environments, providing perceptual information for UAV obstacle avoidance and path planning researches.

## Acknowledgment

This research is supported by the National Research Foundation, Singapore, and the Civil Aviation Authority of Singapore, under the Aviation Transformation Programme. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore and the Civil Aviation Authority of Singapore.

## References

[1] L. Madhuanand, "Monocular depth estimation of uav images using deep learning," Master's thesis, University of Twente, 2020.

[2] Y. Lin, J. Hyyppä, and A. Jaakkola, "Mini-uav-borne lidar for fine-scale mapping," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 3, pp. 426–430, 2010.

[3] E. Cippitelli, F. Fioranelli, E. Gambi, and S. Spinsante, "Radar and rgb-depth sensors for fall detection: A review," *IEEE Sensors Journal*, vol. 17, no. 12, pp. 3585–3604, 2017.

[4] G. T. Leaverton, "Generation drone: The future of utility o&m," in *Electrical Transmission and Substation Structures 2015*, 2015, pp. 190–201.

[5] N. Kong and M. J. Black, "Intrinsic depth: Improving depth transfer with intrinsic images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3514–3522.

[6] A. Rajagopalan, S. Chaudhuri, and U. Mudenagudi, "Depth estimation and image restoration using defocused stereo pairs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 11, pp. 1521–1525, 2004.

[7] M. P. Muresan, M. Negru, and S. Nedevschi, "Improving local stereo algorithms using binary shifted windows, fusion and smoothness constraint," in *2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2015, pp. 179–185.

[8] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia, "Learning monocular depth estimation infusing traditional stereo knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9799–9809.

[9] X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, "Towards real-time monocular depth estimation for robotics: A survey [-5pt]," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[10] G. Chandan, A. Jain, H. Jain *et al.*, "Real time object detection and tracking using deep learning and opencv," in *2018 International Conference on inventive research in computing applications (ICIRCA)*. IEEE, 2018, pp. 1305–1308.

[11] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[12] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[13] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650–2658.

[14] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2024–2039, 2015.

[15] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5162–5170.

[16] H. Florea, V.-C. Miclea, and S. Nedevschi, "Wilduav: Monocular uav dataset for depth estimation tasks," in *2021 IEEE 17th International*

*Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2021, pp. 291–298.

[17] Q. Feng, J. Liu, and J. Gong, "Uav remote sensing for urban vegetation mapping using random forest and texture analysis," *Remote sensing*, vol. 7, no. 1, pp. 1074–1094, 2015.

[18] A. Puri, "A survey of unmanned aerial vehicles (uav) for traffic surveillance," *Department of computer science and engineering, University of South Florida*, pp. 1–29, 2005.

[19] O. Alvear, N. R. Zema, E. Natalizio, and C. T. Calafate, "Using uav-based systems to monitor air pollution in areas with poor accessibility," *Journal of Advanced Transportation*, vol. 2017, 2017.

[20] J. C. Hodgson, S. M. Baylis, R. Mott, A. Herrod, and R. H. Clarke, "Precision wildlife monitoring using unmanned aerial vehicles," *Scientific reports*, vol. 6, no. 1, pp. 1–7, 2016.

[21] S. Ullman, "The interpretation of structure from motion," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 203, no. 1153, pp. 405–426, 1979.

[22] P. Li, D. Farin, R. K. Gunnewiek *et al.*, "On creating depth maps from monoscopic video using structure from motion," in *Proc. of IEEE Workshop on Content Generation and Coding for 3D-television*, 2006, pp. 508–515.

[23] L. Ding and G. Sharma, "Fusing structure from motion and lidar for dense accurate depth map estimation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 1283–1287.

[24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[25] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[26] V.-C. Miclea and S. Nedevschi, "Monocular depth estimation with improved long-range accuracy for uav environment perception," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[27] M. Fonder and M. Van Droogenbroeck, "Mid-air: A multi-modal dataset for extremely low altitude drone flights," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[28] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European conference on computer vision*. Springer, 2016, pp. 740–756.

[29] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.

[30] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.

[31] K. Julian, J. Mern, and R. Tompa, "Uav depth perception from visual images using a deep convolutional neural network," in *Tech. Rep.*, 2017.

[32] A. Marcu, D. Costea, V. Licaret, M. Pîrvu, E. Slusanschi, and M. Leordeanu, "Safeuav: Learning to estimate depth and safe landing areas for uavs from synthetic data," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] Q. Tan, Z. Wang, Y.-S. Ong, and K. H. Low, "Evolutionary optimization-based mission planning for uas traffic management (utm)," in *2019 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2019, pp. 952–958.

[36] M. F. B. Mohamed Salleh, C. Wanchao, Z. Wang, S. Huang, D. Y. Tan, T. Huang, and K. H. Low, "Preliminary concept of adaptive urban airspace management for unmanned aircraft operations," in *2018 AIAA Information Systems-AIAA Infotech@ Aerospace*, 2018, p. 2260.