Ensemble-based Monocular Depth Estimation with Diffusion and Transformer Fusion via Felzenszwalb-Guided Refinement

Tuan-Anh Yang, Minh-Quang Nguyen, Thien-Phuc Tran
VNU-HCM University of Science, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{ytanh21, nmquang21, ttphuc21}@apcs.fitus.edu.vn

Abstract

Monocular depth estimation faces challenges in accurately representing complex scene geometries. We propose a hybrid approach combining Felzenszwalb segmentation and a segment-wise confidence heuristic to improve depth accuracy. Felzenszwalb's algorithm segments images into consistent regions, facilitating depth estimation within structurally coherent areas. A confidence metric based on inverse depth variance allows adaptive selection and refinement of depth estimates at the segment level. This enables effective fusion of diffusion-based (Marigold) and transformer-based (Depth Anything V2) predictions, improving depth consistency and accuracy. Evaluated on the SYNS-Patches dataset, our method demonstrates competitive performance, highlighting the benefits of segment-wise depth fusion.

1. Introduction

Monocular depth estimation is a challenging problem in computer vision that involves predicting scene depth from a single 2D image [2]. Accurate depth estimation is crucial for applications such as robotic perception, autonomous driving, and augmented reality [3]. Traditional methods relying on stereo matching or structured light often require specialized hardware and struggle under varying lighting and textureless surfaces.

Deep learning, particularly Convolutional Neural Networks (CNNs), has improved monocular depth estimation but still faces challenges in capturing global context and complex scene geometry. Transformer-based models have shown promise in modeling long-range dependencies [11], and diffusion-based models have demonstrated strong generative capabilities [18].

We propose a hybrid approach combining Marigold [9], a diffusion-based model, with Depth Anything V2 [17], a transformer-based model. Marigold generates depth estimates, which are refined using the structural insights of Depth Anything V2. We evaluate our method on the SYNS-Patches dataset [13], showing competitive results.

This paper is part of the Monocular Depth Estimation Challenge [12], which encourages novel approaches to improve structural consistency and depth accuracy in complex scenes.

2. Motivation

Monocular depth estimation struggles with accurately capturing complex scene geometries, particularly across regions with varying depth characteristics. We propose a hybrid strategy that combines Felzenszwalb segmentation [6] and a segment-wise confidence heuristic to address this challenge.

Felzenszwalb's algorithm segments images into regions with consistent boundaries, helping to identify areas likely to share similar depth properties. This exploits local structural consistency, as objects or surfaces within a segment should exhibit more uniform depth than regions with sharp discontinuities.

Our confidence heuristic evaluates depth reliability within each segment by computing confidence as the inverse of depth variance. Segments with lower depth variation are treated as more reliable. This enables adaptive depth selection and refinement at the segment level, allowing our method to:

- Select the most reliable depth estimates for each segment.
- Combine the strengths of diffusion-based (Marigold) and transformer-based (Depth Anything V2) models.
- Improve depth fusion beyond simple averaging or maximum likelihood methods.

This segment-level, variance-aware selection mechanism enables more accurate depth fusion and represents a step toward more refined monocular depth estimation.

3. Related Work

3.1. Monocular Depth Estimation

Early depth estimation techniques using handcrafted features and stereo matching were limited in generalizing across scenes [15]. Deep learning revolutionized the field by enabling direct learning of depth cues [11]. Pioneering works like Eigen et al. [5] introduced multi-scale networks, with subsequent research exploring fully convolutional and encoder-decoder architectures to improve prediction accuracy [10].

3.2. Transformer and Diffusion Models

Transformers have expanded from natural language processing to computer vision, demonstrating superior performance in modeling long-range dependencies [4, 16]. Recent models like Depth Anything V2 showcase transformers' ability to capture complex scene geometry [17].

Concurrently, diffusion models have emerged as powerful generative techniques [8]. By iteratively refining predictions through noise reduction, models, such as Marigold, generate detailed depth estimates [9].

3.3. Hybrid Approaches

Our work explores a novel direction of combining diffusion and transformer-based models [7]. By leveraging the generative capabilities of diffusion models and the structural understanding of transformers, we aim to enhance depth estimation accuracy.

4. Methodology

4.1. Problem Formulation

Let $I \in \mathbb{R}^{H \times W \times 3}$ be the input RGB image with height H and width W. We define two depth maps:

- $D_m \in \mathbb{R}^{H \times W}$: Depth map from Marigold
- $D_d \in \mathbb{R}^{H \times W}$: Depth map from Depth Anything V2, converted from disparity to affine-invariant type

4.2. Felzenszwalb Segmentation

We segment the input image I using Felzenszwalb's algorithm to identify consistent regions:

$$S = \text{Felzenszwalb}(I, \text{scale}, \sigma, \text{min\_size}) \tag{1}$$

where $S \in \mathbb{Z}^{H \times W}$ is the segmentation map, with each unique value representing a distinct segment. We use the default parameters: scale $= 200$, $\sigma = 0.8$, and min_size $= 50$.

4.3. Segment-Based Depth Feature Extraction

For each segment $P_i = \{(x, y) \mid S[x, y] = i\}$, we compute:

$$\mu_i = \frac{1}{|P_i|} \sum_{(x,y) \in P_i} D[x, y] \tag{2}$$

$$\sigma_i^2 = \frac{1}{|P_i|} \sum_{(x,y) \in P_i} (D[x, y] - \mu_i)^2 \tag{3}$$

$$c_i = \frac{1}{\sigma_i + \epsilon} \tag{4}$$

We store features for both depth maps as tuples:

$$F_i^m = \{\mu_i^m, (\sigma_i^m)^2, c_i^m\}, \quad F_i^d = \{\mu_i^d, (\sigma_i^d)^2, c_i^d\} \tag{5}$$

$\mu_i^m, (\sigma_i^m)^2, c_i^m$ are mean depth, depth, variance, and confidence extracted from the depth map D_m, and $\mu_i^d, (\sigma_i^d)^2, c_i^d$ are mean depth, depth variance, and confidence extracted from the depth map D_d.

4.4. Adaptive Depth Fusion

Our fusion strategy operates on a segment-wise basis, where T is a confidence threshold, typically we set $T = 1.2$.

Algorithm 1 Adaptive Depth Fusion

for each segment P_i **do**
  Compute mean μ_i and variance σ_i^2 of the depth predictions
  if $\mu_i \approx 1.0$ **and** $\sigma_i^2 < 0.001$ **then**
    $D_{\text{output}}[x, y] \leftarrow 1.0, \quad \forall (x, y) \in P_i$
    continue
  end if
  Compute confidence ratio: $R = \frac{c_i^d}{c_i^m + \epsilon}$
  if $R > T$ **then**
    $D_{\text{output}}[x, y] \leftarrow D_d[x, y], \quad \forall (x, y) \in P_i$
  else
    $D_{\text{output}}[x, y] \leftarrow D_m[x, y], \quad \forall (x, y) \in P_i$
  end if
end for

μ_i and σ_i^2 represent the mean depth and depth variance of any model on the segmentation P_i. The constant $\epsilon > 0$ is set to 10^{-9} to avoid zero division.

Since Marigold's predictions are generally more accurate than Depth Anything V2. We only take the result of Depth Anything V2 if the confidence from Depth Anything V2 is significantly higher than from Marigold (based on threshold T).

5. Results

Table 1 presents a performance comparison of different monocular depth estimation models on the SYNS-Patches dataset, which is a subset of the SYNS dataset [1]. SYNS-Patches consists of 1,656 image and LiDAR pairs extracted

from 92 diverse scenes, including agricultural, natural, residential, industrial, and indoor environments. Each scene contains 18 image patches extracted at 20-degree intervals from a full horizontal rotation at eye level.

Our model demonstrates competitive performance on the SYNS-Patches dataset. Notably, our model achieves an F-score of 14.20 and an edge F-score of 7.81, which are slightly lower than the top-performing models such as PICO-MR [14] (21.07, 8.77) and EVP++ [14] (19.66, 9.02). However, our model shows a relatively higher MAE (4.90) and RMSE (7.96), suggesting that the model's pixel-wise depth estimation may be less accurate compared to other models.

While our model demonstrates reasonable F-score and delta accuracy values, there is room for improvement in terms of absolute depth error and edge accuracy. The higher MAE and RMSE values suggest that refining the model's depth prediction consistency could lead to better overall performance.

6. Conclusion

We proposed a hybrid approach for monocular depth estimation that leverages segment-wise depth fusion to improve accuracy and consistency. Our method combines the generative power of diffusion models (Marigold) with the structural understanding of transformer-based models (Depth Anything V2). By segmenting images using Felzenszwalb's algorithm and applying a variance-based confidence heuristic, our approach adaptively selects the most reliable depth estimates at the segment level. Evaluation on the SYNS-Patches dataset demonstrates competitive performance across multiple metrics, highlighting the effectiveness of segment-wise depth fusion. Future work will explore enhancements to the segmentation strategy and model architecture to further improve depth estimation accuracy and generalization.

References

[1] Wendy Adams, James Elder, Erich Graf, Julian Leyland, Arthur Lugitgheid, and Alexander Muryy. The southampton-york natural scenes (syns) dataset: Statistics of surface attitude. *Scientific Reports*, 6, 2016. 2

[2] Amlaan Bhoi. Monocular depth estimation: A survey, 2019. 1

[3] Xingshuai Dong, Matthew A. Garratt, Sreenatha G. Anavatti, and Hussein A. Abbass. Towards real-time monocular depth estimation for robotics: A survey, 2021. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. 2

[6] Pedro Felzenszwalb and Daniel Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59:167–181, 2004. 1

[7] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation, 2024. 2

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2

[9] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation, 2024. 1, 2, 5

[10] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image, 2014. 2

[11] Uchitha Rajapaksha, Ferdous Sohel, Hamid Laga, Dean Diepeveen, and Mohammed Bennamoun. Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey. *ACM Computing Surveys*, 56(12):1–51, 2024. 1, 2

[12] Jaime Spencer, C. Stella Qian, Chris Russell, Simon Hadfield, Erich Graf, Wendy Adams, Andrew J. Schofield, James Elder, Richard Bowden, Heng Cong, Stefano Mattoccia, Matteo Poggi, Zeeshan Khan Suri, Yang Tang, Fabio Tosi, Hao Wang, Youmin Zhang, Yusheng Zhang, and Chaoqiang Zhao. The monocular depth estimation challenge, 2022. 1

[13] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter, 2022. 1

[14] Jaime Spencer, Fabio Tosi, Matteo Poggi, Ripudaman Singh Arora, Chris Russell, Simon Hadfield, Richard Bowden, GuangYuan Zhou, ZhengXin Li, Qiang Rao, YiPing Bao, Xiao Liu, Dohyeong Kim, Jinseong Kim, Myunghyun Kim, Mykola Lavreniuk, Rui Li, Qing Mao, Jiang Wu, Yu Zhu, Jinqiu Sun, Yanning Zhang, Suraj Patni, Aradhye Agarwal, Chetan Arora, Pihai Sun, Kui Jiang, Gang Wu, Jian Liu, Xianming Liu, Junjun Jiang, Xidan Zhang, Jianing Wei, Fangjun Wang, Zhiming Tan, Jiabao Wang, Albert Luginov, Muhammad Shahzad, Seyed Hosseini, Aleksander Trajcevski, and James H. Elder. The third monocular depth estimation challenge, 2024. 3, 5

[15] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge, 2019. 2

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 2

[17] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. 1, 2, 5

[18] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2024. 1

Ensemble-based Monocular Depth Estimation with Diffusion and Transformer Fusion via Felzenszwalb-Guided Refinement

Supplementary Material

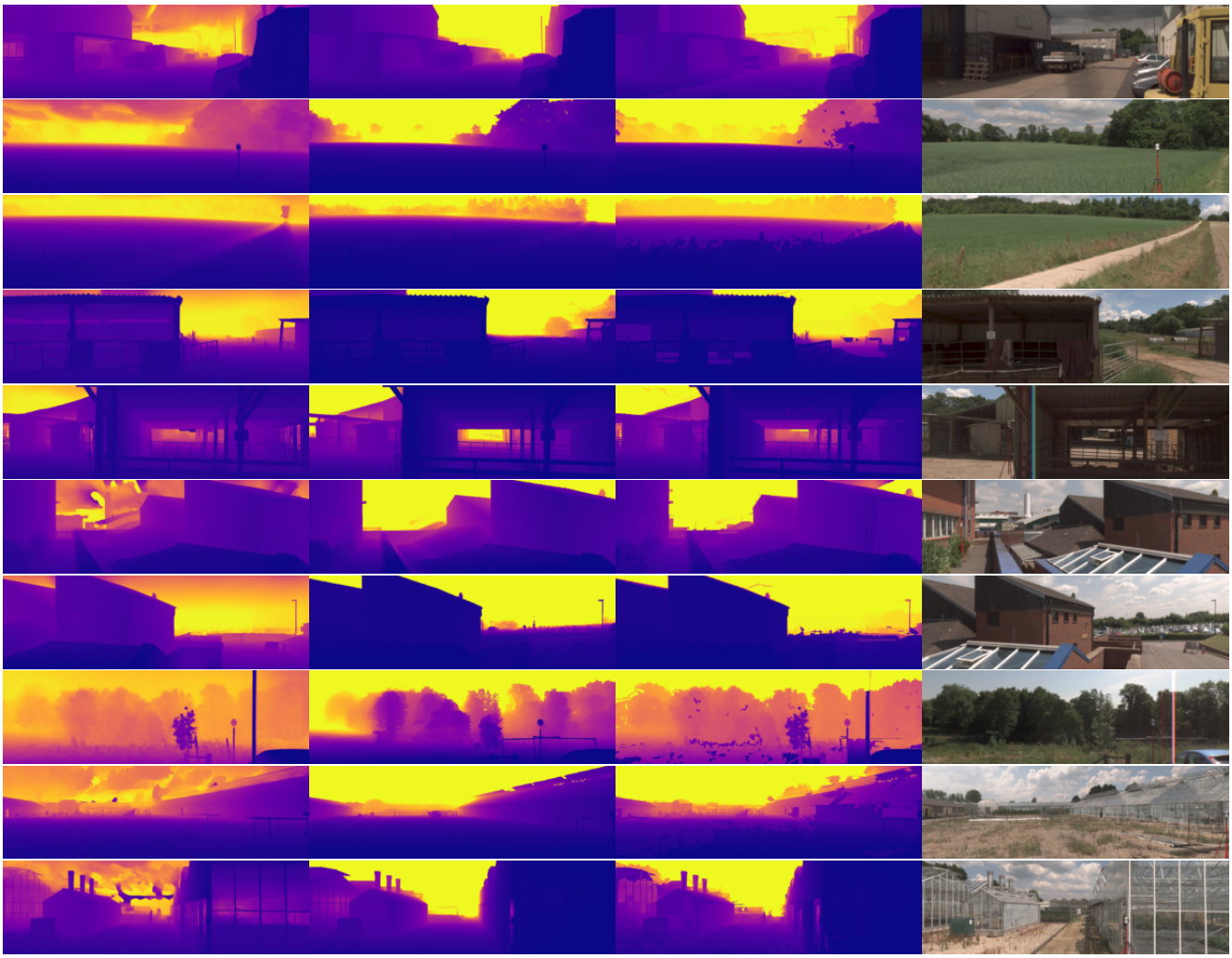

Figure 1. Visualization of the input and output images for each sample. From left to right: (a) Marigold disparity map, (b) Depth Anything V2 disparity map, (c) Refined depth map, and (d) Original image. The shown images correspond to the following IDs: 135, 163, 185, 198, 202, 208, 210, 215, 235, and 243.

Table 1. Performance comparison of different monocular depth estimation models.

Model	F-score	F-score-Edges	MAE	RMSE	AbsRel	EdgeAcc	EdgeComp	Delta1	Delta2	Delta3
Ours	14.20	7.81	4.90	7.96	33.55	3.32	17.66	0.61	0.84	0.92
PICO-MR [14]	21.07	8.77	3.22	5.60	20.33	3.69	15.41	0.7559	0.9125	0.9590
EVP++ [14]	19.66	9.02	3.20	5.49	19.03	2.66	9.28	0.7553	0.9182	0.9661
Marigold [9]	18.64	9.26	3.87	6.49	24.37	2.90	20.09	0.6903	0.8860	0.9453
Depth Anything v2 [17]	14.34	7.94	4.16	7.94	25.48	2.64	30.05	0.6907	0.8849	0.9469
Garg's Baseline [14]	11.38	6.03	4.62	7.58	31.15	4.01	41.24	0.5842	0.8354	0.9251

```python
import numpy as np
import cv2
from skimage.segmentation import felzenszwalb

class SegmentsDepthEnsemble:
    def __init__(self, scale=200, sigma=0.8, min_size=50):
        self.scale, self.sigma, self.min_size = scale, sigma, min_size

    def segment_image(self, image):
        return felzenszwalb(image, scale=self.scale,
                            sigma=self.sigma, min_size=self.min_size)

    def compute_segment_features(self, depth_map, segmentation):
        features = {}
        for i in range(np.max(segmentation) + 1):
            mask = (segmentation == i)
            if mask.sum() > 0:
                features[i] = {
                    'mask': mask,
                    'mean_depth': depth_map[mask].mean(),
                    'std_depth': depth_map[mask].std(),
                    'confidence': 1 / (depth_map[mask].std() + 1e-5),
                    'size': mask.sum()
                }
        return features

    def fuse_depths(self, depth1, depth2, segments1, segments2,
                    conf_threshold=1.2):
        fused_depth = depth1.copy()
        for i in segments1:
            if i in segments2:
                s1, s2 = segments1[i], segments2[i]
                mask = s1['mask']
                if (s1['mean_depth'] > 0.999 and s1['std_depth'] < 0.001) or \
                (s2['mean_depth'] > 0.999 and s2['std_depth'] < 0.001):
                    fused_depth[mask] = 1.0
                    continue
                conf_ratio = s2['confidence'] / (s1['confidence'] + 1e-5)
                if conf_ratio > conf_threshold:
                    fused_depth[mask] = depth2[mask]
        return fused_depth

    def ensemble_depth_maps(self, depth1, depth2, image):
        segmentation = self.segment_image(image)
        seg1 = self.compute_segment_features(depth1, segmentation)
        seg2 = self.compute_segment_features(depth2, segmentation)
        return self.fuse_depths(depth1, depth2, seg1, seg2)
```

Figure 2. Python implementation of `SegmentsDepthEnsemble` class for depth map fusion.