Vol.2025-CVIM-241 No.12
2025/3/18

# 単眼深度推定モデルのテスト時最適化による Depth from Focus

小橋口 純[1,a)]　藤村 友貴[1]　北野 和哉[1]　舩冨 卓哉[1]　向川 康博[1]

**概要**：Depth from Focus (DFF) は，カメラのフォーカス距離の変化を利用して深度を推定する技術である．深層学習を用いた DFF はスケールが正確な深度を推定できるが，学習データが少なく精度が低いという課題がある．一方で，近年 Depth Anything 等の，大規模なデータセットで学習された単眼深度推定の基盤モデルが提案されている．これらの基盤モデルで推定された深度は，定性的には優れているがスケールは不定である．したがって，本研究では DFF と Depth Anything を組み合わせ，スケールが正確かつ定性的にも優れた深度推定手法を提案する．具体的には，推論時に DFF の出力を利用し，Depth Anything のパラメータをシーンごとに最適化する．実画像のデータセットで評価を行い，提案手法を DFF の従来手法と比較し，有効性を確認した．

**キーワード**：Depth from Focus，単眼深度推定，Test-time Optimization

## 1. はじめに

深度推定は，画像を入力としてカメラから物体までの距離を推定する技術であり，自動運転，AR/VR，ロボットビジョンなどに応用されている．深度推定の中でも Depth from Focus (DFF) は，カメラのフォーカス距離の変化を利用して深度を推定する技術である．入力はフォーカルスタックと呼ばれる，フォーカス距離を変えて撮影した複数の画像の組である．出力は，入力画像の画素ごとに深度を計算した深度マップである．

LiDAR などのセンサは高精度に深度を計測することができるが，画像による深度推定の方が高い解像度の深度マップが得られる．そのため，細部の形状を計測したい場合は，DFF を含めた画像による深度推定が適している．また，ステレオ法など複数の視点から撮影した画像を用いる方法に比べ，DFF は 1 台のカメラで深度を推定できる．そのため，小型化やコスト削減を実現できる可能性がある．

これまでに深層学習を用いた DFF の手法が提案されている．DFV [5] では，フォーカルスタック内の画素ごとに最も焦点が合ったフレームを推定することにより，スケールが正確な絶対深度を推定している．しかし，DFV は学習に用いたデータが少ないため，精度や汎化性能には限界がある．

一方で，近年コンピュータビジョンの分野では，大規模なデータセットで学習された基盤モデルの事前知識を，様々なタスクに活用する研究が進められている．深度推定においては，1 枚の画像を入力として深度マップを推定する単眼深度推定の基盤モデルが提案されている．

Depth Anything [6], [7] は半教師あり学習を応用し，深度の正解がラベル付けされたデータセットに加え，ラベル付けされていない大規模なデータセットも用いて学習を行う．これにより，定性的に優れた深度マップを推定することができる．しかし，深度のスケールを無視した損失関数を用いることにより，スケールの異なる複数のデータセットで学習を行っているため，推定される深度はスケールの情報を含まない相対深度となる．

そこで本研究では，DFF と単眼深度推定を組み合わせることにより，定性的に優れた絶対深度を推定する手法を提案する．具体的には，推論時に DFF の出力を利用し，単眼深度推定モデルのパラメータをシーンごとに最適化（テスト時最適化）する．単眼深度推定モデルをテスト時最適化する試みに，SfM-TTR [2] が挙げられる．SfM-TTR では，Structure from Motion の推定結果を用いて単眼深度推定モデルを最適化し，深度の精度を向上させることに成功している．本研究では，単眼深度推定モデルが出力する相対深度を，焦点が合う距離に近づけるように最適化することで，スケールが正確な絶対深度に変換する手法を提案する．

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[a)] kohashiguchi.jun.kh0@naist.ac.jp

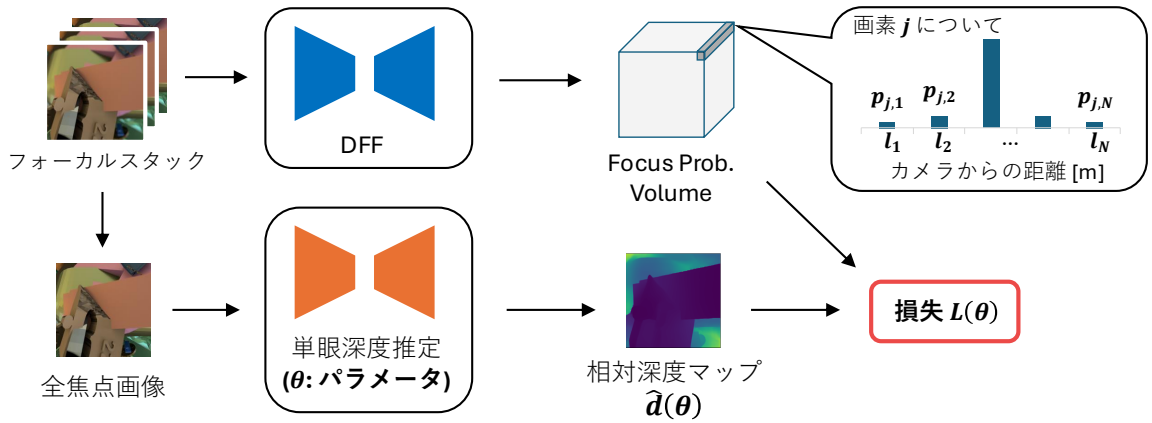


**図 1** 提案手法の全体像．DFF の出力を利用して単眼深度推定モデルのパラメータをテスト時最適化することにより，相対深度マップを絶対深度に変換する．

## 2. 提案手法

図 1 に提案手法の全体像を示す．提案手法では，推論時に単眼深度推定モデルのパラメータをシーンごとに最適化する．DFF と単眼深度推定のそれぞれの出力を得た後，それらを用いて損失を計算し最適化を行う．

### 2.1 DFF

DFF では，フォーカルスタック $x \in \mathbb{R}^{N\times H\times W}$ を入力とし，入力画像の画素ごとに深度を推定する．ここで，画像サイズが $H \times W$，フレーム数が $N$ である．また，フレームを $i$，画素を $j$ で表し，$i$ 番目のフレームの画素 $j$ の値を $x_{j,i}$ と表す．

DFV [5] はフォーカルスタックを入力とし，focus probability volume $p$ を出力する．$p$ は $p \in \mathbb{R}^{N\times H\times W}$ であり，$p_j \in \mathbb{R}^N$ はある画素 $j$ においてどのフレームに最も焦点があっているかの確率を表す．例えば，$p_{j,i}$ は画素 $j$ の $i$ 番目のフレームに最も焦点が合っている確率である．このような確率分布を推定することにより，フォーカルスタックのフレームの間隔よりも細かい精度で，焦点が合う距離を推定することができる．具体的には，次の式に従って focus probability volume から絶対深度 $d_j$ が計算される．

$$d_j = \sum_i p_{j,i} l_i \tag{1}$$

$l_i$ はフォーカルスタックの $i$ 番目のフレームを撮影した際のフォーカス距離を表す．

### 2.2 単眼深度推定

単眼深度推定では 1 枚の画像から相対深度を推定する．これに対し，本研究の入力はフォーカルスタックであり複数枚の画像が存在する．しかしながら，これらの画像は全てぼけを含んでいるため，どの画像を入力しても全焦点画像を入力した場合に対して精度が低下することが分かった．ここで，全焦点画像とは，全ての画素で焦点が合っており，ぼけのない画像である．そこで，本研究ではまず最初に，focus probability volume $p$ とフォーカルスタック $x$ を用いて，全焦点画像 $x'$ を作成する．式 (1) と同様に，全焦点画像は以下の式に基づいて作成する．

$$x'_j = \sum_i p_{j,i} x_{j,i} \tag{2}$$

作成した全焦点画像を単眼深度推定モデルに入力し，相対深度 $\hat{d}$ を取得する．単眼深度推定モデルには，Depth Anything [6], [7] など大規模なデータセットで学習された基盤モデルを用いる．そうすることで，様々なシーンの相対深度に関する事前知識を活用した，定性的に優れた深度推定結果を得ることができる．

### 2.3 テスト時最適化

提案手法では，DFF の出力を利用し，単眼深度推定の相対深度 $\hat{d}$ を絶対深度に変換する．単純なアプローチとして，式 (3) のように最小二乗法で求めたパラメータ $a$,$b$ を用いて，深度マップ全体を線形に変換するアプローチが考えられる．

$$\min_{a,b} \|a\hat{d} + b - d\|^2 \tag{3}$$

しかし，線形変換は表現力が低く，絶対深度としての精度には限界がある．

そこで，単眼深度推定モデルのパラメータ $\theta$ を推論時に最適化することにより，非線形に変換するアプローチを提案する．具体的には式 (4) のように，DFF が出力する focus probability volume $p$ を用いて，相対深度 $\hat{d}(\theta)$ を画素ごとに焦点が合う距離に近づける．

$$\min_{\theta} \sum_j \sum_i p_{j,i} (l_i - \hat{d}_j(\theta))^2 \tag{4}$$

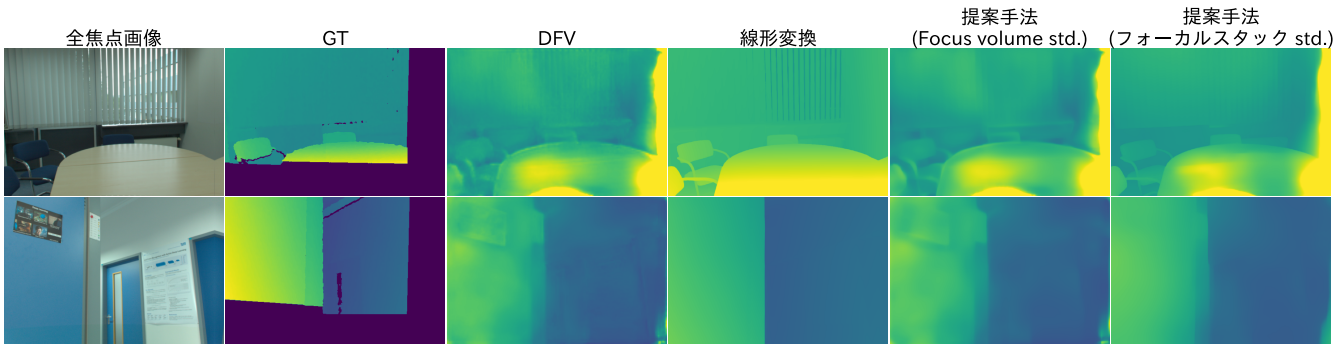


**図 2** 深度推定結果．GT は正解の深度マップである．

**表 1** 定量評価による提案手法と従来手法の比較．指標ごとに最も精度の良いものを太字で示している．実験の結果，提案手法の精度は従来手法（DFV）を上回ることが確認された．

| Method | MSE ↓ | RMS ↓ | Abs.rel. ↓ | $\delta_1$ ↑ | $\delta_2$ ↑ | $\delta_3$ ↑ |
|---|---|---|---|---|---|---|
| DFV | 0.000570 | 0.0213 | 0.174 | 76.7 | 94.2 | 98.1 |
| 線形変換 | 0.000719 | 0.0227 | 0.202 | 70.8 | 93.3 | 97.4 |
| 提案手法 (Focus volume std.) | 0.000570 | 0.0213 | 0.174 | 76.7 | 94.2 | 98.1 |
| 提案手法 (フォーカルスタック std.) | **0.000564** | **0.0212** | **0.172** | **77.3** | **94.3** | **98.2** |

これにより，相対深度 $\hat{d}(\theta)$ を非線形に変換し，スケールが正確な絶対深度を得る．

### 2.4 DFF の信頼度

しかしながら，式 (4) のように最適化すると，$\hat{d}_j(\theta)$ を式 (1) で得られた $d_j$ に近づけるだけであり，DFF の精度を越えることができない（付録参照）．そこで提案手法では，DFF の出力を信頼できる度合い $m_j$ を画素ごとに計算し，それを用いて重み付けをして最適化を行う．

$$\min_{\theta} \sum_j m_j \sum_i p_{j,i}(l_i - \hat{d}_j(\theta))^2 \tag{5}$$

ここで，信頼度 $m_j$ は 0 から 1 の値をとり，大きいほど DFF の出力が信頼できることを表す．信頼度を用いることで，信頼度が高い領域では DFF の出力に近づけ，それ以外の領域では単眼深度推定モデルが持つ事前知識により深度を推定する．本研究では，$m_j$ の計算方法を 2 つ検討した．

**focus probability volume の標準偏差**

2.1 節で述べたように，DFV などの DFF 手法では画素ごとに焦点の合う距離を表す確率分布を計算する．しかし，テクスチャのない領域などの焦点が合うフレームの推定が難しい画素や，そもそも焦点の合うフレームが存在しない画素では，確率分布が広く分散すると考えられる．そこで，画素ごとに確率分布の標準偏差を計算し，DFF の信頼度として用いる．

$$\mu_j = \sum_i p_{j,i} l_i \tag{6}$$

$$\sigma_j = \sum_i p_{j,i}(l_i - \mu_j)^2 \tag{7}$$

$$m_j = -\exp(\sigma_j) \tag{8}$$

**フォーカルスタックの標準偏差**

DFF はカメラのフォーカス距離の変化を利用し，最も焦点が合う距離を推定する．しかし，先ほど述べたようにテクスチャのない領域ではカメラのフォーカス距離を変えても画素値の変化が小さいため，DFF による推定が難しい．そこで，次式のようにフォーカルスタックのフレーム方向に画素値の標準偏差を計算し，DFF の信頼度として用いる．

$$\mu_j = \frac{1}{N}\sum_i x_{j,i} \tag{9}$$

$$\sigma_j = \frac{1}{N}\sum_i (x_{j,i} - \mu_j)^2 \tag{10}$$

$$m_j = 1 - \exp(\sigma_j) \tag{11}$$

## 3. 実験

実画像のデータセットを用いて定性的・定量的な評価を行い，提案手法を DFF の従来手法である DFV [5] と比較した．また，DFF の信頼度の計算方法の違いによる，精度の変化も調べた．

### 3.1 データセット

DDFF-12 [1] は，ライトフィールドカメラを用いて撮影された，実画像のフォーカルスタックのデータセットである．RGB-D センサを用いて取得された正しい深度が提供

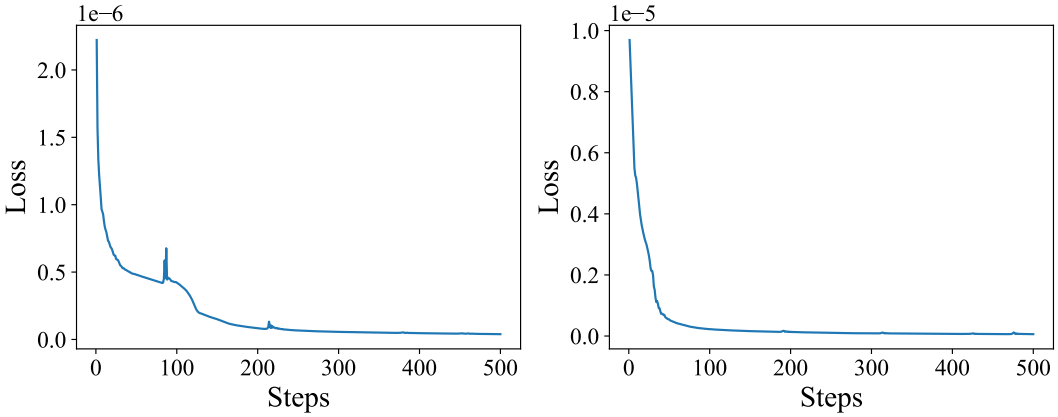


**図 3** 損失のプロット．横軸は反復回数，縦軸は損失を表す．

されており，深層学習を用いた DFF 手法の評価に用いられる．データセットは屋内のシーンで構成されており，DFF にとって難易度の高い，壁や机などテクスチャのない物体も多く含まれている．

データセットの各サンプルは，フォーカルスタックと深度マップのペアで構成される．12 個のシーンのうち，6 個のシーンはそれぞれ 100 個のサンプルを含み，残りの 6 個のシーンはそれぞれ 20 個のサンプルを含む．実験では，100 個のサンプルを含むシーンのうち，2 つのシーンを用いて評価を行った．

### 3.2 実験条件

提案手法は DFF と単眼深度推定を組み合わせて構成されるが，DFF の手法には DFV，単眼深度推定の手法には Depth Anything V2 [7] を用いた．Depth Anything V2 はエンコーダとして DINOv2 [3] を使用し，デコーダとして DPT head [4] を用いている．本実験では SfM-TTR [2] を参考にし，学習済みの Depth Anything V2 のパラメータを初期値とし，エンコーダのパラメータのみ最適化を行った．

提案手法の実装には PyTorch を用いた．また，最適化手法は Adam を用い，学習率を $10^{-5}$，反復回数を 500 回とした．

### 3.3 評価指標

推定された絶対深度の精度を評価する指標として， [1] で使用されているものと同じ評価指標を利用した．MSE，RMS，Absolute relative は，正しい深度との誤差を計算する指標であり，小さい方が精度が良いことを表す．正解率 $\delta_1$，$\delta_2$，$\delta_3$ は，正しい深度と推定された深度の比が閾値よりも小さい画素の割合を計算する指標であり，大きい方が精度が良いことを表す．

### 3.4 実験結果

#### 3.4.1 定性評価

図 2 に，DFV，線形変換，提案手法の深度推定結果を示す．線形変換は，Depth Anything V2 の出力を式 (3) で求めたパラメータを用いて線形に変換したものである．また，提案手法には DFF の信頼度として focus probability volume の標準偏差を用いる場合と，フォーカルスタックの標準偏差を用いる場合の 2 通りがある．

フォーカルスタックの標準偏差を DFF の信頼度として用いた場合，図 2 の 1 段目のサンプルでは，提案手法は DFV に比べ椅子の細部の深度を正しく推定している．2 段目のサンプルでは，提案手法は DFV に比べ平らな壁の深度を正しく推定している．また，提案手法は線形変換よりも，深度の値が Ground truth に近くなっている．一方で，focus probability volume の標準偏差を信頼度として用いた場合は，提案手法の出力は DFV の出力に類似したものになった．

これらのことから，フォーカルスタックの標準偏差を信頼度として用いると，提案手法は Depth Anything V2 の定性的に優れた深度マップを維持しつつ，定量的にも優れた深度を推定できることが分かった．

#### 3.4.2 定量評価

表 1 に，定量評価の結果を示す．表の数値は，DDFF-12 の 200 サンプルのそれぞれで深度推定を行い算出した精度の平均値である．定量評価の結果，フォーカルスタックの標準偏差を DFF の信頼度として用いた場合，提案手法の精度が DFV の精度を上回ることが確認された．さらに，線形変換は DFV よりも精度が低く，提案手法の非線形な変換が定量評価において優れていることが確認できた．

また，focus probability volume の標準偏差を DFF の信頼度として用いる場合に比べ，フォーカルスタックの標準偏差を信頼度として用いる方が精度が高くなることが分かった．

#### 3.4.3 計算時間

定性評価に用いた 2 つのサンプルについて，最適化にかかる時間を計測した．最適化には NVIDIA RTX 6000 Ada（48GB）を用いた．提案手法の最適化時間は 1 つ目のサンプルでは 4 分 41 秒であり，2 つ目のサンプルでは 4 分 39 秒だった．

また，損失のプロットを図 3 に示す．横軸は反復回数，縦軸は損失である．実験では反復回数は 500 回としたが，それよりも早い段階で収束する場合もあり，実用上はより短い時間で最適化できる可能性もある．

## 4. まとめ

本研究では，DFF の出力を利用し単眼深度推定モデルを推論時に最適化する新しい DFF の手法を提案した．DFF のフォーカス情報を利用し，単眼深度推定モデルが出力する深度を焦点が合う距離に近づけるように最適化することで，スケールが正確な絶対深度を推定した．定性，定量評価により，提案手法の精度は従来手法を上回ることが確認された．

**謝辞**

本研究は JSPS 科研費 JP22K17911 の助成を受けたものである．

### 参考文献

[1] Hazirbas, C., Soyer, S. G., Staab, M. C., Leal-Taixé, L. and Cremers, D.: Deep Depth From Focus, *ACCV* (2018).
[2] Izquierdo, S. and Civera, J.: SfM-TTR: Using Structure from Motion for Test-Time Refinement of Single-View Depth Networks, *CVPR*, pp. 21466–21476 (2023).
[3] Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A. and Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision, *TMLR* (2024).
[4] Ranftl, R., Bochkovskiy, A. and Koltun, V.: Vision Transformers for Dense Prediction, *ICCV*, pp. 12179–12188 (2021).
[5] Yang, F., Huang, X. and Zhou, Z.: Deep Depth From Focus With Differential Focus Volume, *CVPR*, pp. 12642–12651 (2022).
[6] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J. and Zhao, H.: Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data, *CVPR*, pp. 10371–10381 (2024).
[7] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J. and Zhao, H.: Depth Anything V2, *NeurIPS* (2024).

## 付　　録

$$\min_{\theta} \sum_{j} \sum_{i} p_{j,i}(l_i - \hat{d}_j(\theta))^2 \tag{A.1}$$

$$\Leftrightarrow \min_{\theta} \sum_{j} \sum_{i} p_{j,i}(l_i^2 - 2l_i\hat{d}_j(\theta) + \hat{d}_j(\theta)^2) \tag{A.2}$$

$$\Leftrightarrow \min_{\theta} \sum_{j} \sum_{i} -2p_{j,i}l_i\hat{d}_j(\theta) + p_{j,i}\hat{d}_j(\theta)^2 \tag{A.3}$$

$$\Leftrightarrow \min_{\theta} \sum_{j} -2d_j\hat{d}_j(\theta) + \hat{d}_j(\theta)^2 \tag{A.4}$$

$$\Leftrightarrow \min_{\theta} \sum_{j} (d_j - \hat{d}_j(\theta))^2 \tag{A.5}$$

したがって，式 (4) の最適化は $\hat{d}_j(\theta)$ を式 (1) で得られた $d_j$ に近づけることと等価である．