# 胸装着型カメラによる三次元姿勢推定の深層学習用合成データ生成手法の提案

早川 恭平 * Dong-Hyun Hwang† Chen-Chieh Liao* 小池 英樹 *

**概要.** 深層学習において実データの収集が困難な場合, 合成データを使用することが一般的であるが, 実データとのドメインギャップが課題となる. 一人称視点の魚眼画像から三次元姿勢推定を行う MonoEye[2] では, 同様の問題により推論精度が低下する. 本論文では, 画像生成モデルである Stable Diffusion[6] を使用し, 既存の合成画像より現実の画像ドメインに近い画像を生成する手法を提案する. 生成画像に対して定性および定量評価を行い, 生成画像がドメインギャップに起因する推論精度低下を緩和する可能性を示した.

## 1 はじめに

深層学習において実データの収集が困難な場合, 合成データを使用することが一般的であるが, 実データと合成データとのドメインギャップが課題となる.

魚眼カメラを用いた一人称視点の三次元姿勢推定手法 [2, 3, 5, 7, 8, 9, 10, 12] においても, その多くが合成画像の使用に依存している. 特に, Hwang らの MonoEye[2] は, 胸部に装着した魚眼カメラからの画像を元に装着者の三次元姿勢を推定する手法であり, 他の手法とはセットアップが異なるため, 実データの収集が一段と困難である.

本論文では, 画像生成モデルである Stable Diffusion[6] と画像の構図の詳細な制御を行う ControlNet[11] を使用して, MonoEye の合成画像より現実の画像ドメインに近い画像を生成する手法を提案する. また, 提案手法で生成された画像で学習した深層学習モデルにおける, 現実の画像に対する推論精度について定量的に検証を行う.

## 2 提案手法

### 2.1 画像生成モデル用データセットの作成

Stable Diffusion および ControlNet の学習のためのデータセットを作成する. Stable Diffusion の学習には画像およびそれに対応するキャプションが必要である. ControlNet の学習には, これらに加えて構図を制御する条件画像が必要である. 本論文では, 条件画像として OpenPose[1] を模倣した人物の姿勢画像を採用する.

学習用の画像は MonoEye データセット [2] を参考に Unity を使用して作成する (図 1(a)). キャプションは, 画像の作成に使用する人物モデルと背景画像のキャプションを組み合わせる. 姿勢画像は, 自己中心視点からの魚眼画像内の人物の姿勢を再現するために, Unity 上で人物モデルに対応するボーンとリグを作成し, 仮想魚眼カメラで撮影することで作成する (図 1(b)). それぞれのデータを 10,000 個ずつ作成し, 学習用データセットとする.

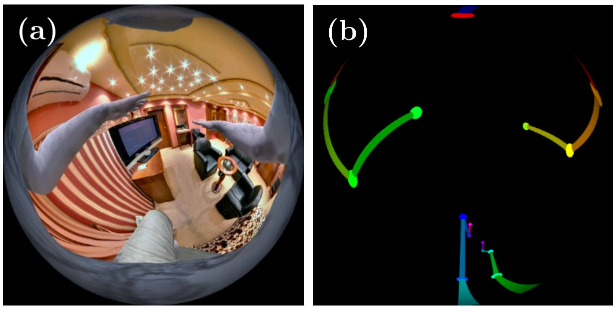


**図 1. 画像生成モデル用データセットの例.(a)Unity で作成した魚眼画像.(b) 対応する姿勢画像. (a) に対するキャプションは, ”bedroom, shorts and a t-shirt, skin color is white”である.**

加えて, Stable Diffusion で現実の画像ドメインに近似した画像を生成するために, 現実の画像を用いたファインチューニング用のデータセットを作成する.2 名の被験者を対象として, 複数の環境下で魚眼画像を撮影し, 適切なキャプションを作成する. この画像とキャプションを 24 組作成する.

### 2.2 画像生成モデルの学習

画像生成モデルの学習は, 次の三段階に分けて実施する.

1. 合成魚眼画像とキャプションを使用して, Stable Diffusion をファインチューニングする.
2. 1. のモデルに対して, 合成魚眼画像, キャプションおよび姿勢画像を使用し, ControlNet を学習する.
3. 現実の魚眼画像とキャプションを使用して, 1. のモデルをさらにファインチューニングする.


* 東京工業大学
† NAVER Cloud

**表 1. 各動作カテゴリと全体の平均の PA-PMJPE の計算結果 (単位: mm).**

| データセット | 部位 | Walk | Crouching | Sitting | boxing | Crawling | Stretching | Waiving | 全体 |
|---|---|---|---|---|---|---|---|---|---|
| MonoEye | 上半身 | 103.05 | 226.45 | 213.21 | 122.64 | 313.9 | 151.6 | 136.36 | 183.99 |
| | 下半身 | 76.98 | 212.21 | 193.81 | 101.2 | 277.75 | 173.05 | 69.17 | 162.35 |
| | 全身 | **89.15** | 218.85 | 202.87 | 111.2 | 294.62 | 163.04 | 100.53 | 172.45 |
| 提案手法 | 上半身 | 106.51 | 203.83 | 205.97 | 101.63 | 310.29 | 99.39 | 96.73 | 163.25 |
| | 下半身 | 90.07 | 192.81 | 189.3 | 80.71 | 276.64 | 130.39 | 53.00 | 148.71 |
| | 全身 | 97.74 | **197.95** | **197.08** | **90.48** | **292.34** | **115.92** | **73.41** | **155.49** |

### 2.3 画像生成

現実の魚眼画像でファインチューニングしたモデルと ControlNet を使用し, キャプションと姿勢画像を入力とし, 画像の生成を行う. 姿勢画像は Unity で作成されるため, 画像上での関節座標と三次元空間上での関節座標が関連付けられている. したがって, 生成画像において姿勢画像の各座標が正解データとして関連付けられる.

画像はランダムに 10,000 枚生成し, 生成時間は, NVIDIA GeForce RTX 3090 一台で約 3.4 時間であった. 生成画像の例を図 2 に示す.

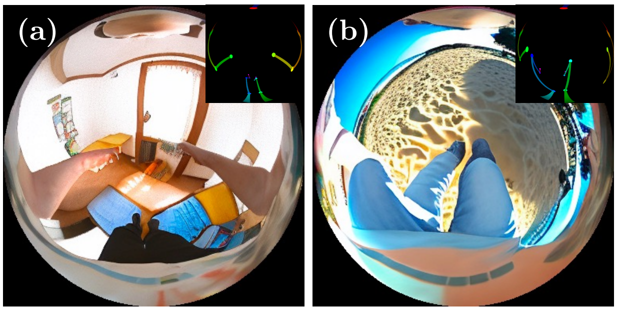


**図 2. 提案手法で生成された画像. 画像右上は生成に使用した姿勢画像.(a) 品質の高い画像の例.(b) 品質の低い画像の例.**

## 3 実験

### 3.1 定性評価

合成魚眼画像 (図 1(a)) と生成魚眼画像 (図 2(a)) を視覚的に比較し, 評価を行った.

### 3.2 定量評価

合成画像と生成画像のぞれぞれで学習した MonoEye の BodyPoseNet[2] について, 現実の画像のテストデータセットに対する推論精度を比較した.

合成画像での学習は新たに行わず, MonoEye の学習済みモデルを使用した. また, 生成画像での学習は, MonoEye の学習済みモデルのパラメータで初期化し, ファインチューニングを行った.

テストデータセットは, 学習データとは異なる 2 名の被験者の動作カテゴリ（表 1）に対し, マーカーレスモーションキャプチャ[1]を適用し, 作成した.

推論精度の指標には, Procrustes 解析 [4] を適用した MPJPE(PA-MPJPE) を採用した. 各モデルに対して PA-MPJPE を算出し, 各動作カテゴリと全体の平均推論精度を表 1 に示す.

## 4 考察

### 4.1 定性評価

図 1(a) と図 2(a) を比較すると, 前者では人の肌の色合いが灰色の傾向が見受けられるのに対し, 後者ではベージュ色に近似しており, 現実のテクスチャにより近似していると考えられる. また, 図 2(a) より, 背景や人物に魚眼レンズ特有の歪みが適切に再現されていることが確認できる. これらより, 合成画像と比べ, 生成画像が現実の画像ドメインにより近似していると考えられる. しかし, 図 2(b) に示すように, 生成画像の中には条件画像の期待する姿勢とは異なる品質の低い画像も存在するため, 生成された全ての画像が現実の画像ドメインに近似しているとは言えない.

### 4.2 定量評価

表 1 より, MonoEye の学習済みモデルと比較し, 生成画像で学習したモデルの全体の PA-MPJPE の平均値が低いことがわかる. この結果は, 提案手法を用いて生成された画像が, 現実のデータに対する推論精度低下を緩和する可能性を示唆している.

しかし, Walk の動作カテゴリの PA-MPJPE の平均値は, MonoEye の学習済みモデルが低いことがわかる. これは, 生成画像の中に図 2(b) のような期待する姿勢とは異なる品質の低い画像が存在することが一因であると考えられる. また, 学習済みモデルと比較し, 学習に使用した生成画像の数が少ないことも一因であると考えられる.

## 5 まとめ

Stable Diffusion と ControlNet を使用し, 既存の合成魚眼画像より現実の画像ドメインに近い魚眼画像を生成する手法を提案した. 提案手法で生成した画像を定性および定量評価し, ドメインギャップに起因する推論精度低下を緩和する可能性を示した.

[1] https://qualisys.archivetips.com

## 謝辞

本論文は JST CREST JPMJCR17A3 の支援を受けている.

## 参考文献

[1] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021.

[2] D.-H. Hwang, K. Aso, Y. Yuan, K. Kitani, and H. Koike. MonoEye: Multimodal Human Motion Capture System Using A Single Ultra-Wide Fisheye Camera. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, p. 98–111, New York, NY, USA, 2020. Association for Computing Machinery.

[3] H. Jiang and V. K. Ithapu. Egocentric Pose Estimation from Human Vision Span. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10986–10994, 2021.

[4] D. G. Kendall. A Survey of the Statistical Theory of Shape. *Statistical Science*, 4(2):87–99, 1989.

[5] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt. EgoCap: Egocentric Marker-Less Motion Capture with Two Fisheye Cameras. *ACM Trans. Graph.*, 35(6), dec 2016.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022.

[7] D. Tome, P. Peluse, L. Agapito, and H. Badino. xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7727–7737, 2019.

[8] J. Wang, L. Liu, W. Xu, K. Sarkar, D. Luvizon, and C. Theobalt. Estimating Egocentric 3D Human Pose in the Wild with External Weak Supervision. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13147–13156, 2022.

[9] J. Wang, L. Liu, W. Xu, K. Sarkar, and C. Theobalt. Estimating Egocentric 3D Human Pose in Global Space. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11480–11489, 2021.

[10] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, P. Fua, H.-P. Seidel, and C. Theobalt. $Mo^2Cap^2$ : Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019.

[11] L. Zhang, A. Rao, and M. Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models, 2023.

[12] Y. Zhang, S. You, and T. Gevers. Automatic Calibration of the Fisheye Camera for Egocentric 3D Human Pose Estimation from a Single Image. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1771–1780, 2021.

## 未来ビジョン

本論文で画像生成モデルの学習に使用した画像は 10,000 枚と, 現代の画像生成モデルで使用されるデータセットと比較して非常に少ない. これは我々の持つ計算資源の制約に起因する. したがって, データ量を増やせば, 生成画像の品質向上が期待できる. 特に, ファインチューニングに使用した現実の画像は 2 人分と限られていた. より多くの人のデータを取得し, 画像のバリエーションを増やすことで, さらなる品質向上が見込まれる.

また, 我々の研究の最終的な目的は, 胸装着型カメラを用いた三次元姿勢推定の現実のデータに対する推論精度の向上である. 使用している姿勢推定モデルのアーキテクチャは約 4 年前のものであるため, 最新の技術を導入することでさらなる精度向上が期待される. 具体的には, CNN ベースから Transformer ベースのアーキテクチャへの変更や, 時系列データの導入などが考慮される.

実際の画像における推論精度の向上は, 胸装着型カメラを用いた三次元姿勢推定のアプリケーション領域を広げる可能性がある. 我々は, この研究を通じて, ヒューマンコンピュータインタラクションの分野に新たな寄与を目指す.