対応点推定におけるインスタンス情報を用いた誤マッチングの抑制

Xinmeng Yu[1,2,a)]　薗頭 元春[,1]　北野 和哉[1]　藤村 友貴[1]　向川 康博[1]　川西 康友[3,2,1,b)]

概要

本研究では，インスタンス情報を用いて対応点推定の信頼性を向上させる手法を提案する．LoFTR などのセミデンス対応点推定手法は，低テクスチャ領域でも多くの対応点を推定できる．しかし，室内環境に複数の類似物体が存在する場合，画像特徴のみでは正しい対応関係を判断することが困難であり，異なる物体の対応点を誤って推定することがある．このような誤マッチングは，対応点推定結果の信頼性を低下させ，相対姿勢推定などの後続処理にも影響する．この課題に対処するため，本研究では Instance-LoFTR を提案する．提案手法では，まずインスタンスセグメンテーションにより画像中のインスタンスマスクを生成し，LoFTR のマッチング過程にインスタンスレベルの制約を導入する．

1. はじめに

対応点推定は，二枚の画像間でシーン中の同じ点に対応する画像上の点の位置を求める処理であり，コンピュータビジョンにおける基本的な技術の一つである．画像間の対応関係を求めることで，カメラの相対的な位置や姿勢の推定，3 次元形状の復元，画像検索，物体認識など，さまざまな処理に利用できる [1]．そのため，対応点推定の精度は，後続の処理全体の信頼性に大きく関わる．さまざまな環境に対して正確で安定した対応点推定を実現することは，現在でも重要な課題である．

従来の対応点推定では，まず画像中から特徴点を検出し，その周辺の特徴量を記述したうえで，画像間の特徴量を比較して対応関係を求める手法が広く用いられてきた．代表的な手法である SIFT [3] は，スケールや回転の変化に対して比較的安定した局所特徴量を抽出できるため，長く利用されてきた．また，近年では SuperPoint [4] のように，深層学習を用いて特徴点の検出と特徴量の記述を同時に行う手法も提案されている．さらに，SuperGlue [5] や LightGlue [6] のように，検出された特徴点間の対応関係をニューラルネットワークにより推定する手法も発展している．これらの特徴点ベースの手法は，対応候補を特徴点に限定するため計算量を抑えやすく，多くの場面で有効である．しかし，壁，床，机の表面のような低テクスチャ領域では，安定した特徴点を十分に検出できない場合がある．また，特徴点が少ない場合には，対応点の分布が偏り，画像全体の対応関係を十分に表現できないことがある．

図 1　類似物体間に生じる誤マッチングの例．LoFTR を用いた対応点推定実験により，類似物体間に誤マッチングが生じることを確認した．図中，緑線は正しい対応を示すのに対し，赤線は誤った対応を示す．対応図の左上には，全対応点数とそのうちの正しい対応点数を示す．

このような問題に対して，近年では特徴点検出に依存しない detector-free な画像マッチング手法が注目されている．LoFTR [1] は，二枚の画像から抽出した特徴マップに対して Transformer の self-attention と cross-attention を適用し，画像間の文脈を考慮した特徴表現を得る．また，粗い対応を求めた後に細かい対応位置を補正する coarse-to-fine の構成により，密な対応点を生成する．ASpanFormer [7] は，適応的な attention span を用いることで，画像間の大域的な文脈と局所的な対応関係を考慮する．これらの detector-free 手法により，特徴点ベースの手法では対応が

1　奈良先端科学技術大学院大学
2　理化学研究所
3　立命館大学
a)　xinmeng.yu@riken.jp
b)　kwns@fc.ritsumei.ac.jp

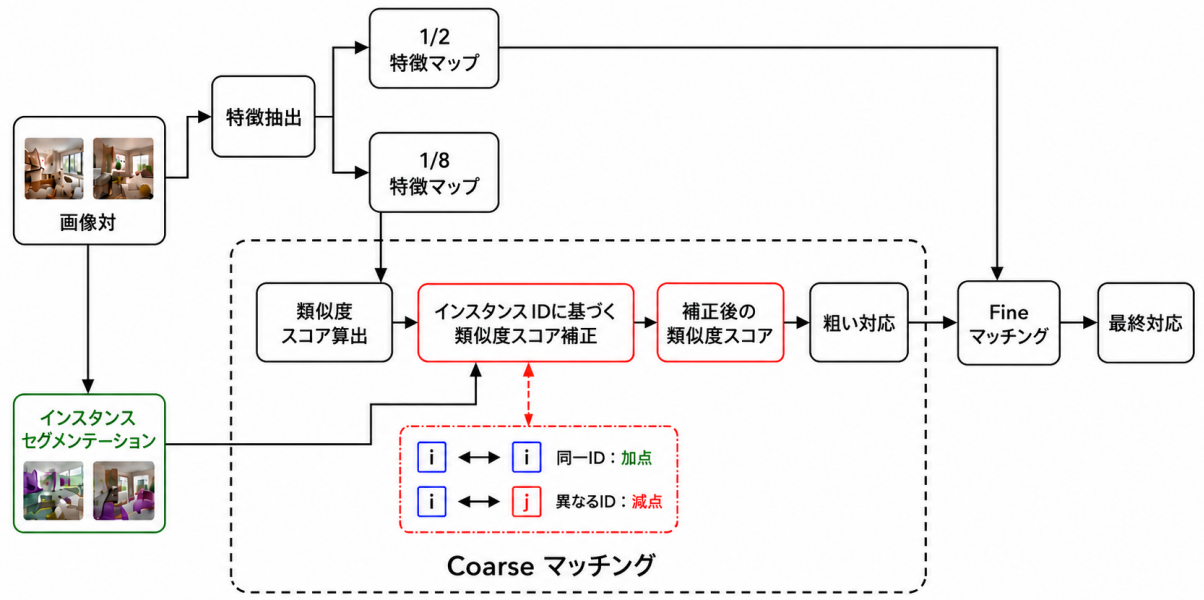

図 2 Instance-LoFTR の概要．インスタンスセグメンテーションにより得られたインスタンス情報を用いて，coarse マッチングにおける類似度スコアを補正する．

困難な低テクスチャ領域においても対応点を得ることができる．しかし，これらの手法は主に画像特徴に基づいて対応関係を判断するため，画像中に外観の似た物体が複数存在する場合には，異なる物体間で誤マッチングが生じる可能性がある．図 1 に，類似物体間に生じる誤マッチングの例を示す．

本研究では，類似物体間で生じる誤マッチングを抑制するために，インスタンス情報を対応点推定に導入する Instance-LoFTR を提案する．インスタンス情報は，同じカテゴリに属する物体であっても個々の物体を区別できるため，外観が類似した物体が複数存在する場面において，画像特徴のみでは判断しにくい対応関係を補助できると考えられる．LoFTR は，低テクスチャ領域でも比較的密な対応点を生成できる手法であるため，本研究の対応点推定手法として用いる．本研究ではインスタンスセグメンテーションにより得られるマスクを利用し，LoFTR のマッチング過程にインスタンスレベルの情報を反映させる．これにより，低テクスチャ領域に強い LoFTR の利点を保ちながら，類似物体間の誤マッチングを減らし，より信頼性の高い対応点推定の実現を目指す．ScanNet [8] データセットを用いた評価により，Instance-LoFTR が，類似物体間に生じる誤マッチングを抑制できることを確認する．

2. 提案手法

本研究では，入力画像対と各画像に対応するインスタンスマスクを用い，対応候補の位置が属するインスタンス ID に基づいてマッチングスコアを補正する Instance-LoFTR を提案する．図 2 に概要を示す．

2.1 Instance-LoFTR

Instance-LoFTR は入力画像に対してインスタンス情報を付与する処理と，その情報を類似度スコアに反映する処理から構成される．

インスタンス情報の付与 入力画像対の各画像に対してインスタンスセグメンテーションを行い，各画素にインスタンス ID を割り当てる．対応点推定では，形状や色などの外観が類似した物体が複数存在する場合，それらの物体が誤って対応付けられる可能性がある．また，カテゴリ情報のみを用いる場合，同じ種類で外観も類似した物体は同一カテゴリとして扱われるため，個々の物体を区別することができない．そこで本手法では，インスタンスセグメンテーションにより，各画素がどの物体に属しているかを表すインスタンス ID を取得する．これにより，同じカテゴリに属する物体であっても，それぞれを異なる物体として扱うことができる．

インスタンス情報の反映 画像特徴に基づいて類似度スコアを算出した後，対応候補となる二点がそれぞれどのインスタンスに属しているかを確認する．画像特徴のみを用いる場合，外観が類似した異なる物体間でも高い相似度スコアが得られる可能性があるため，そのままマッチングを行うと誤マッチングが生じる．そこで本手法では，インスタンス ID を類似度スコアの補正に用いる．具体的には，対応候補の二点が同じ物体に対応すると考えられる場合にはその対応を相対的に選ばれやすくし，異なる物体に対応すると考えられる場合にはその対応を選ばれにくくする．

2.2 ボーナスとペナルティのルール

LoFTR では，二枚の画像から抽出された特徴マップ間

の類似度スコアを計算する．この類似度スコアに対して，対応候補となる二点のインスタンス ID に応じた補正を行う．補正前の類似度スコアを S_{ij}，補正後の類似度を $\widehat{S}_{ij}$，対応候補となる二点のインスタンス ID をそれぞれ I_i，I_j とする．また，背景に属する画素の ID を 0 とする．補正後の類似度スコアは以下のように定義する．

$$\widehat{S}_{ij} = \begin{cases} S_{ij} + \alpha, & I_i = I_j,\ I_i \neq 0,\ I_j \neq 0 \\ S_{ij}, & I_i = 0,\ I_j = 0 \\ S_{ij} - \beta, & \text{otherwise} \end{cases} \tag{1}$$

ここで，α は同一物体内の対応に与えるボーナス，β は異なる物体間，または物体と背景の間の対応に与えるペナルティである．対応候補の二点が同じインスタンス ID を持ち，かつ背景ではない場合，それらは同一物体に属する点であると判断し，スコアにボーナスを加える．一方，二点のインスタンス ID が異なる場合，または一方が物体で他方が背景である場合には，異なる領域間の対応であると判断し，スコアにペナルティを与える．

背景同士の対応については，補正せず，元の類似度スコアをそのまま用いる．これは，インスタンスセグメンテーションが画像中のすべての物体を完全に検出できるとは限らず，背景として扱われた領域の中にもマッチングに有効な情報が含まれる可能性があるためである．背景同士の対応にペナルティを与えると，このような有効な対応まで抑制してしまう可能性がある．そのため，本手法では背景同士の対応を補正対象から除外する．

3. 実験および考察

3.1 実験データ

本研究では，実験データとして ScanNet データセット [8] を用いる．ScanNet は，室内環境を対象とした大規模 RGB-D 動画像データセットであり，1,500 以上の scan から得られた約 250 万の画像を含む．各 scan には，RGB 画像と深度画像に加えて，三次元カメラ姿勢，表面再構成結果，インスタンスレベルのセマンティックセグメンテーションが付与されている．本研究では，ScanNet に含まれる RGB 画像を用いて画像対を作成し，画像マッチングの実験に使用する．

3.2 実験概要

本研究では，LoFTR [1] と同様に，対応点推定の性能を相対姿勢推定により評価する．2 枚の画像間で得られたマッチング点を用いてカメラ間の相対回転および相対移動を推定し，推定された相対姿勢と正解姿勢との誤差を算出する．

実験では，インスタンスマスクの精度が提案手法に与える影響を確認するため，2 つの設定を用いる．第 1 の設定では，YOLOv8 [9] によりテストセットの画像をセグメンテーションし，得られたインスタンスマスクを用いてマッチングを行う．予備的な確認では，補正前の類似度スコアはおおむね 50 以下であったため，元の類似度スコアを大きく変化させない補正値として，ボーナスおよびペナルティはともに 1.5 とする．

第 2 の設定では，理想的なインスタンス分割が得られた場合の上限的な性能を確認するため，ScanNet に含まれる ground truth マスクを用いる．ただし，ScanNet のテストデータには ground truth マスクが含まれていないため，本設定では ScanNet の訓練データを用いる．この設定では，インスタンス ID の信頼性が高いことを前提とし，ボーナスを 0.5，ペナルティを 100,000 とする．これにより，異なる ID を持つ物体間の対応をほぼ除外し，正確なインスタンス情報が画像マッチングに与える影響を評価する．

各設定では，それぞれ 1,500 組の画像対を用いて画像マッチングを行い，相対姿勢推定の誤差がそれぞれ 5°，10°，20° 以下となる範囲で算出した Area Under the Curve（AUC）により性能を比較する．

表 1 YOLOv8 により生成したインスタンスマスクを用いた定量結果.

AUC	LoFTR	Instance-LoFTR
@5°	22.13	**22.85**
@10°	40.79	**41.25**
@20°	57.75	**58.17**

表 2 Ground truth を用いた定量結果.

AUC	LoFTR	Instance-LoFTR
@5°	**25.99**	25.58
@10°	44.51	**45.03**
@20°	62.29	**62.85**

3.3 実験結果

表 1 は YOLOv8 により生成したインスタンスマスクを用いた定量結果，図 3 は 定性結果である．YOLOv8 により生成したインスタンスマスクを用いた場合，Instance-LoFTR はすべてのしきい値において LoFTR を上回った．この結果から，推定されたインスタンスマスクを用いた状況において，インスタンス情報に基づくスコア補正が相対姿勢推定の精度向上に有効であることが確認できる．

表 2 は ground truth マスクを用いた定量結果，図 4 は定性結果である．Ground truth マスクを用いた場合，Instance-LoFTR は AUC@10° および AUC@20° において LoFTR を上回った．一方で，AUC@5° では LoFTR の方が高い値を示した．これは，ground truth マスクを用いた実験では ScanNet の訓練データを用いており，LoFTR がすでに同データセットを学習しているため，LoFTR による対応点推定が有利に働いた可能性がある．また，ScanNet の ground truth マスクは 3D インスタンスマスクを 2D

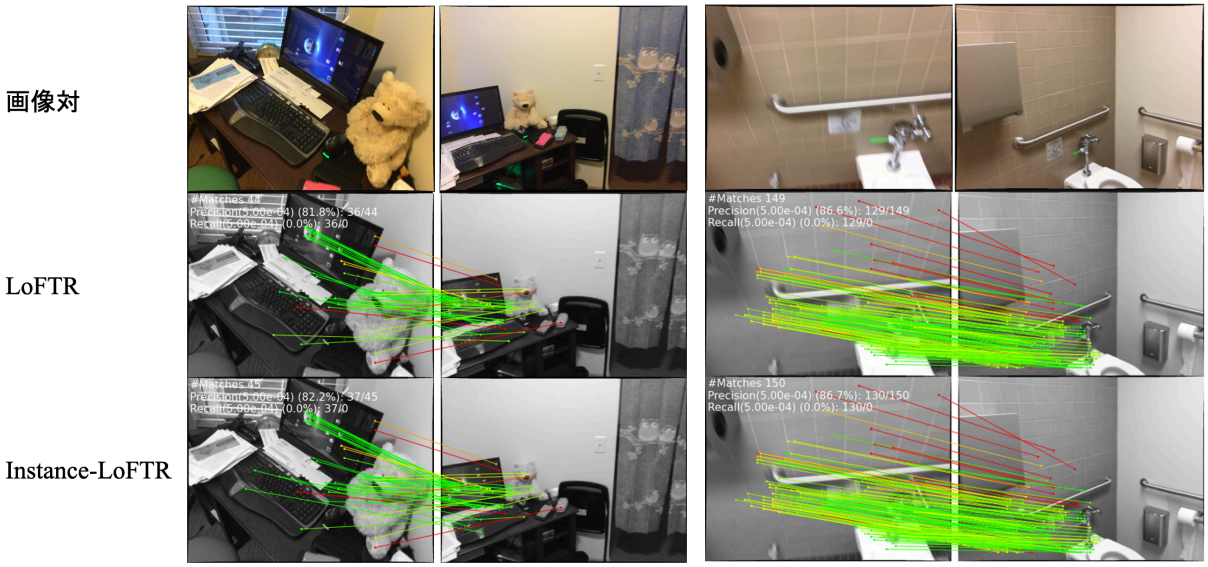

図 3 YOLOv8 により生成したインスタンスマスクを用いた定性結果．緑線は正しい対応を示すのに対し，赤線は誤った対応を示す．対応図の左上には，全対応点数とそのうちの正しい対応点数を示す．

図 4 Ground truth マスクを用いた定性結果．緑線は正しい対応を示すのに対し，赤線は誤った対応を示す．対応図の左上には，全対応点数とそのうちの正しい対応点数を示す．

画像へ投影して得られるため，境界が粗く，5° のような厳しいしきい値では対応点推定の精度に逆に悪影響を及ぼした可能性がある．しかし，AUC@10° および AUC@20° では Instance-LoFTR が LoFTR を上回っており，ground truth マスクを用いた場合でも，インスタンス情報に基づくスコア補正が相対姿勢推定の精度向上に有効であることが確認できる．また図 4 より，提案手法により異なる物体間の誤マッチングが大きく抑制され，対応点推定の正確性が向上していることが分かる．

4. おわりに

本研究では，インスタンス情報を用いて対応点推定における誤マッチングを抑制する Instance-LoFTR を提案した．提案手法では，LoFTR の類似度スコアをインスタンス ID に基づいて補正し，外観が類似した異なる物体間の対応を抑制した．ScanNet データセットを用いた実験により，YOLOv8 により生成したインスタンスマスクを用いた場合に LoFTR より高い AUC を示し，ground truth マスクを用いた場合にも多くのしきい値で性能向上を確認した．一方で，対応の正確性の向上が相対姿勢推定の精度向上に十分につながらない場合も見られたため，今後は相対姿勢推定に有効な対応点の選択方法を検討する．

参考文献

[1] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, “LoFTR: Detector-Free Local Feature Matching with Transformers,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8922–8931, 2021.

[2] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou, “Efficient LoFTR: Semi-Dense Local Feature Matching with Sparse-Like Speed,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21666–21675, 2024.

[3] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, Vol. 60, No. 2, pp. 91–110, 2004.

[4] D. DeTone, T. Malisiewicz, and A. Rabinovich, “SuperPoint: Self-Supervised Interest Point Detection and Description,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 224–236, 2018.

[5] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, “SuperGlue: Learning Feature Matching with Graph Neural Networks,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4938–4947, 2020.

[6] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, “LightGlue: Local Feature Matching at Light Speed,” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17627–17638, 2023.

[7] H. Chen, Z. Luo, L. Zhou, Y. Tian, M. Zhen, T. Fang, D. McKinnon, Y. Tsin, and L. Quan, “ASpanFormer: Detector-Free Image Matching with Adaptive Span Transformer,” Proceedings of the European Conference on Computer Vision, pp. 20–36, 2022.

[8] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2432–2443, 2017.

[9] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLOv8,’’ GitHub repository, 2023. Available: https://github.com/ultralytics/ultralytics

[10] G. Potje, F. Cadar, A. Araujo, R. Martins, and E. R. Nascimento, “XFeat: Accelerated Features for Lightweight Image Matching,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2682–2691, 2024.