第 30 回日本バーチャルリアリティ学会大会論文集（2025 年 9 月）

映像から影と物体との対応を認識するモデル

川田裕貴 [1)]，阿部亨 [2)]
Yuki KAWATA, Toru ABE

1) 東北大学大学院情報科学研究科（〒 980-8578 宮城県仙台市青葉区荒巻字青葉 6-3, yuki.kawata.r6@dc.tohoku.ac.jp）
2) 東北大学サイバーサイエンスセンター（〒 980-8578 宮城県仙台市青葉区荒巻字青葉 6-3, beto@tohoku.ac.jp）

概要: AR 環境で自然な影を描写するには，物体とその影の関係を正確に捉える技術が必要となる．本稿では，物体とその影を区別せずに検出する既存手法を用い，その中で明度や彩度といった色の情報の変化や，画像の再構成を行った際に生じる誤差を用いて影と物体の領域を区別して認識する処理を導入し，検出された領域の配置や位置関係をもとに影と物体の対応を認識する手法を提案する．

キーワード： 画像処理，拡張現実，影

1. はじめに

近年，仮想物体を現実世界に自然に融合させる技術が進展し，AR 分野への応用が広がっている．その中でも影の描写は非常に重要な要素であり，影が適切に生成されることで，仮想物体の存在感や実在感が大きく向上する．[1,2]．しかし現状の AR アプリケーションでは，影が物体の輪郭のみを参考に一様に付加されることが多く，実際の光源の方向や強さに応じた自然な影表現は実現されていない [1,2]．

仮想物体に自然で整合的な影を生成するためには，まず画像中の実際の影と物体を識別し，それぞれの位置や形状を把握する必要がある．仮想物体に影を付加する際には，背景画像に写り込む影及びそれに対応する物体の領域を検出し，これらの位置関係や形状・色といった特徴を基に影の生成が行われるためである [1]．とりわけ重要なのは,「物体に映る影」ではなく,「物体が環境に落とす影」の検出であり，物体とその影の対応関係を認識することが求められる．[3–5] そこで，本研究では「物体が環境に落とす影」のことを対象とする影とする．物体以外の背景に生じる影であ

本研究では，最終的な影の生成を行うモデルそのものには焦点を当てず，その前段階に位置付けられる「影と物体の識別および対応関係の推定」の処理に着目する．参照対象となる物体および対応する影のマスクを高精度に抽出することは，後段階の影生成処理における自然さや整合性の向上に寄与すると期待される．

このような観点から，画像中の影と物体を個別に識別し，それらの対応関係までを含めて推定する技術が注目されている．代表的な手法として Single-Stage Instance Shadow detection network（SSIS）[4] は，影と物体の一対一の対応関係を明示的に学習する構成を持つ．一方で，このような対応関係付きの検出を行うには，影と物体をペアとした教師データが必要である．また，教師データを用意する都合上，想定する環境を広げるほど，大量のデータが必要である，という問題も抱えている．

近年は，このような負担を軽減すべく，自己教師型学習による注目領域の検出手法も登場している．Sauvalle らの研究 [6] では，視覚的な一貫性やパターン性から外れる領域（＝影や物体など）を，教師データなしに検出する，再構成誤差に基づく領域の検出手法が提案されている．

この仕組みの中核となるのは,「背景だけを復元する処理」である．ここで言う「背景だけ」とは，その 1 枚の画像において，周囲のパターンや構造から予測可能な領域のことである．たとえば一様な床の模様は，周囲の見えているほかの床の模様と連続しているため，そこにどのようなパターンがあるかを推測できる．このように，同一画像内にある周囲情報から復元可能な部分を「背景」として再構成している．一方で，影や物体は「その部分だけが急に暗くなっている」「周囲とつながりがない模様で覆われている」といった性質を持つため，周囲の情報だけから推測することが難しく，再構成結果では正しく復元されない．

結果として，影や物体のあった領域では，入力画像と再構成画像との間に差（＝誤差）が発生する．このように,「この画像においてどこが推測可能か／不可能か」という判断が，影や物体などの特徴的な領域とそれ以外の背景を分ける基準になっている．そのため，教師データがなくとも，再構成画像と入力画像の差分をとることで，影や物体などの特徴的な領域が検出できる．

さらに，Cross-Attentional Dual Decoder Network（CADDN）[5] では，画像中に存在する影領域を，Sauvalle らの研究 [6] をもとにした再構成誤差に基づいた検出と，従来の教師あり学習による検出を併用することで，SSIS に代表されるような完全な教師あり学習に比べて必要なアノテーション量を削減しながら，多様な環境への対応も見せている．しかし，CADDN では検出は影のみに限定されており，影と物体の検出と，その対応関係の推定までは行われていない．

本研究では，こうした影生成に先立つ処理の信頼性を向

上させることを目的とし，Sauvalle らの「自己教師型学習を用いた特徴的な領域の検出」と，CADDN の「再構成誤差をベースとした影検出構造」を踏襲しつつ，教師データを必要とせずに影と物体の領域を検出し，それを明度・彩度の変化に基づいて影と物体を分離し，その対応関係を明示的に推定する新たな手法を提案する．

2. 関連研究

2.1 教師あり学習による影-物体対応関係の検出

影と物体の対応関係を明示的に扱う代表的な手法として，SSIS [4] が挙げられる．SSIS は，影と物体を個別に検出した上で，それらの対応関係（どの影がどの物体に対応するか）を同時に学習する構成を持つ．この構成により，影と物体の対応関係を高い精度で推定できることが特徴である．

一方で，このような対応関係の学習を行うには，画像ごとに物体・影の領域に対する教師データを作成する必要がある．そのため，精度向上や対応力の向上には相応のデータ量が求められる．

2.2 自己教師型学習の注目領域検出

教師データなしで画像中の物体や影の領域を検出する手法として，Sauvalle らの研究 [6] がある．この手法では，画像の再構成を行うことで得られる再構成誤差に基づいて，入力画像中の背景から逸脱する領域を検出する．こうした領域は，視覚的に特徴的な要素（＝物体や影）である可能性が高いとされる．

このアプローチでは，学習時に教師データを用意する必要がなく，完全に自己教師型学習の構成で動作可能である．ただし，検出されるのはあくまで前景と背景の違いに基づく注目領域であり，それぞれの領域が何を意味するのか（影か物体か），あるいはどのような対応関係にあるかといった情報までは得られない．

2.3 自己教師型学習と教師あり学習を併用した影検出

CADDN [5] は，Sauvalle らのように特徴的な領域を自己教師型学習を用いて検出する点では共通しているが，影の教師データを用いた教師あり学習を併用することで，注目領域を「影」として明確に定義・学習できる構造を持つ．ただし，影と物体の同時検出や，影と物体の対応関係（どの物体に属する影か）を推定する構成は持っていない．

2.4 本研究の位置づけ

現時点では，教師データを用いずに影と物体の対応関係までを自動的に推定する手法は確立されておらず，本研究はそのような手法の構築に向けた一つの試みである．

3. 提案手法

本研究では，影と物体の対応関係を教師データなしで推定することを目指し，自己教師型学習の構成による 3 ステップの処理を提案する．

- **ステップ 1：領域検出**: 入力画像の再構成誤差に基づいて特徴的な領域（影・物体の候補）を検出
- **ステップ 2：領域分離**: 色の特徴を利用して影と物体を分離
- **ステップ 3：対応関係推定**: 分離された領域間の空間的な関係をもとに，対応関係を推定

3.1 ステップ 1：領域検出

このステップでは，Sauvalle らの研究 [6] や CADDN [5] で用いられた方法と同様に，再構成誤差を利用して入力画像中の特徴的な領域を検出する．

高さ H，幅 W の RGB 画像を，画像中の RGB 値，輝度，彩度，境界線の長さや方向といった幾何学的特徴をベクトルとして表現した入力画像 I を，U-Net をベースとした自己再構成ネットワーク $R(\cdot)$ に入力し，再構成された出力画像 $\hat{I} = R(I)$ を得る．

次に，入力画像と再構成画像の各チャンネル（画像中の R, G, B 成分）における画素値の差の絶対値を平均することで再構成誤差マップ E を求める．画像中の位置 (x, y) における誤差 $E(x, y)$ は，以下の式で与えられる：

$$E(x,y) = \frac{1}{3} \sum_{c \in R,G,B} \left| I_c(x,y) - \hat{I}_c(x,y) \right|. \tag{1}$$

ここで，$I_c(x, y)$ および $\hat{I}_c(x, y)$ はそれぞれ，元画像と再構成画像におけるチャンネル c（R, G, B）の位置 (x, y) における画素値を表す．

得られた誤差マップ E に対し，全体の平均誤差 μ_E および標準偏差 σ_E を計算し，これらを用いてしきい値 τ を以下のように設定する：

$$\tau = \mu_E + \lambda \cdot \sigma_E. \tag{2}$$

ここで，λ は再構成誤差に対する感度を調整するためのハイパーパラメータである．

このしきい値 τ を用いることで，次式により再構成誤差が大きい箇所を影・物体候補領域として抽出したマスクを作成する：

$$M(x,y) \quad = \quad \begin{cases} 1 & E(x,y) > \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

3.2 ステップ 2：領域分離

ステップ 1 では再構成誤差により，影と物体を含む領域が検出されるが，この時点では両者の区別はなされていない．

既存手法では，HSV 色空間に基づく単純なしきい値処理により影と物体の分離を試みる [3] が，この方法では濃色物体や逆光などの条件で誤認が発生しやすい．本研究ではこの問題に対処するため，色・構造・再構成特性の 3 要素に基づく自己教師型の分離処理を行う．

（1）色空間に基づく初期クラスタリング

まず，ステップ 1 のマスク領域 $M(x, y)$ に対して画像を HSV 色空間に変換し，明度 $V(x, y)$ と彩度 $S(x, y)$ を用いて影の初期候補マスク $M_s^{(0)}$ を作成する：

$$M_s^{(0)}(x,y) \ = \ \begin{cases} 1 & V(x,y) < \mu_V - \alpha \cdot \sigma_V \\ & \quad \text{and} \ \ S(x,y) < \mu_S - \beta \cdot \sigma_S, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

ここで μ_V, σ_V はマスク領域内の明度平均・標準偏差，μ_S, σ_S は彩度の統計量，α, β は感度調整パラメータである．この処理は [3] で述べられている手法と同様のものであり，低明度かつ低彩度の画素が影候補として抽出されるが，濃色物体や照明ムラとの誤区別は困難なため，さらなる分離処理が必要となる．

（2）構造的特徴に基づく相互補正

初期候補マスク $M_s^{(0)}$（影）および $M_o^{(0)} = M - M_s^{(0)}$（物体）に対し，両者の構造的・視覚的な違いに着目し，マスク間の整合性を高める自己教師的な補正処理を行う．以下に示す各種スカラー特徴を (x, y) に対して計算し，それらの組み合わせにより影スコアおよび物体スコアを求める：

- **局所明度・彩度平均**：対象位置周辺の 5×5 パッチ内における明度 $\bar{V}(x, y)$，彩度 $\bar{S}(x, y)$ の平均
- **境界エッジ強度** $E(x, y)$：局所エッジ勾配の大きさ
- **マスク間距離** $D(x, y)$：影マスク内の画素と物体マスクの最近傍距離，またはその逆

上記の特徴を用いて，各画素 (x, y) における影の信頼度スコア $S_{\text{shadow}}(x, y)$ を以下で評価する：

$$\begin{aligned} S_{\text{shadow}}(x, y) = \omega_1 \cdot \left(1 - \text{CosSim}\left(\bar{V}_s(x, y), \bar{V}_o(x, y)\right)\right) \\ + \omega_2 \cdot (1 - \text{EdgeClosedness}(x, y)) \\ + \omega_3 \cdot \text{DistToObject}(x, y). \end{aligned} \tag{5}$$

ここで，$\text{CosSim}(\cdot, \cdot)$ は 2 つの明度ベクトルのコサイン類似度，$\text{EdgeClosedness}(x, y)$ は影候補領域の境界がどれだけ閉じているかを評価する関数，$\text{DistToObject}(x, y)$ は前述のマスク間距離 $D(x, y)$ に基づいて正規化された距離関数である．

同様に，物体候補画素に対する信頼度スコア $S_{\text{object}}(x, y)$ は以下で評価する：

$$\begin{aligned} S_{\text{object}}(x, y) = \theta_1 \cdot \text{ColorContrastToBackground}(x, y) \\ + \theta_2 \cdot \text{ContourSharpness}(x, y) \\ + \theta_3 \cdot \text{DistanceToShadow}(x, y). \end{aligned} \tag{6}$$

ここで，$\text{ColorContrastToBackground}(x, y)$ は背景領域との色のコントラストを計算する指標，$\text{ContourSharpness}(x, y)$ は画素周辺の輪郭がどれだけ明瞭であるかを評価する指標，$\text{DistanceToShadow}(x, y)$ は物体から最も近い影領域までの距離に基づく関数である．

なお，ω_i および θ_i（$1 \le i \le 3$）は各スコア項目に対応する重み係数であり，特徴の重要度を調整する役割を持つ．

各スコアに対して閾値を設定し，信頼度が低い画素をマスクから除外または再割り当てすることで，中間マスク $M_s^{(1)}$，$M_o^{(1)}$ を得る．この補正処理は，マスク間の整合性が収束するまで繰り返し適用される．

（3）自己再構成による整合性チェック

さらに，$M_s^{(1)}$ に基づいて，この時点で予測された影領域を黒塗りした画像を再構成ネットワーク $R(\cdot)$ に再入力し，予測された影領域を黒塗りした画像 $\tilde{I} = R(I \odot (1 - M_s^{(1)}))$ を得る．この予測された影領域を黒塗りした結果と元画像 I との差を以下で求める：

$$S_{\text{recons}}(x, y) = \left| I(x, y) - \tilde{I}(x, y) \right|. \tag{7}$$

ここで，$S_{\text{recons}}(x, y)$ は予測された影領域を黒塗りした結果と元画像との誤差に基づく影の信頼度スコアであり，小さいほど「影としてマスクしても自然に補完される」すなわち影である可能性が高いと判断できる指標である．

このスコアに対して閾値を設定することで，最終的な影マスク $M_s^{(2)}$ を確定し，残りの領域 $M_o = M - M_s^{(2)}$ を物体の領域とみなす．

3.3 ステップ 3：対応関係推定

ステップ 2 により得られた影マスク M_s および物体マスク M_o に対して，教師なしで影と物体の対応関係を推定する．既存の SSIS モデルでは，主に空間的近接性や重なり度に基づいて対応ペアを構築していたが，学習時に正解ラベルを必要とするため，完全な自己教師型とは言えなかった．

本研究では，影領域の集合 $\{S_i\}_{i=1}^{N}$ および物体領域の集合 $\{O_j\}_{j=1}^{M}$ の各ペア (S_i, O_j) に対して以下のようなスコアを求めることで対応関係を推定する．

（1）オフセットベクトル

各影領域 S_i および物体領域 O_j に対して，その重心 $\mathbf{c}_{S_i}, \mathbf{c}_{O_j}$ と重心間のオフセットベクトル $\mathbf{v}_{i,j} = \mathbf{c}_{S_i} - \mathbf{c}_{O_j}$ を求める．

これら S_i, O_j および $\mathbf{v}_{i,j}$ に基づき，影 S_i と物体 O_j のペアに対する対応信頼度を以下の $d_{i,j}, o_{i,j}$ により評価する：

- 距離スコア $d_{i,j} = \|\mathbf{v}_{i,j}\|$：影と物体の重心間距離（ユークリッド距離）
- 重なりスコア $o_{i,j} = \text{IoU}(S_i, O_j)$：2 つの領域の交差面積と合計面積の比率（Intersection over Union）

これらを統合し，最終的な対応スコア $r_{i,j}$ を次式で求める：

$$r_{i,j} = w_d \cdot \exp(-\gamma \cdot d_{i,j}) + w_o \cdot o_{i,j} \tag{8}$$

ここで w_d, w_o は各項に対する重み係数，γ は距離に対する減衰係数である．このようにして得られるスコア $r_{i,j}$ をもとに，影 S_i に対して最もスコアが高い物体 O_j を対応先として選択する．ただし，対応スコアが閾値 τ_r を下回る場合は「対応なし」とみなし，除外することで誤ペアリングを抑制する．

（2）対応構造への一貫性損失導入

さらに，予測された対応構造の妥当性を高めるために，一貫性損失 $\mathcal{L}_{\text{consistency}}$ を導入する．

各対応ペア (S_i, O_j) に対して，以下の 2 つの観点から損失を評価する：

- 明度差損失：領域内の平均明度の差 $\Delta V = |\bar{V}_{O_j} - \bar{V}_{S_i}|$
- 境界滑らかさ損失：2 つのマスク境界上における勾配不一致度 $\text{EdgeDiscrepancy}(S_i, O_j)$

図 1: 入力画像に対する処理結果の成功例．左：入力画像，中央：検出結果，右：分離結果

図 2: 入力画像に対する処理結果の失敗例．左：入力画像，中央：検出結果，右：分離結果

ここで，$\bar{V}_{S_i}$ および $\bar{V}_{O_j}$ はそれぞれの領域内の画素の平均明度を表す．最終的な損失関数 $L_{\text{consistency}}$ は以下のように表される：

$$\mathcal{L}_{\text{consistency}} = \sum_{(i,j)} I[r_{i,j} > \tau_r] \cdot \big(\lambda_1 \cdot |\bar{V}_{O_j} - \bar{V}_{S_i}| + \lambda_2 \cdot \text{EdgeDiscrepancy}(S_i, O_j)\big). \tag{9}$$

ここで，$I[\cdot]$ は条件成立時に 1 を返すインジケータ関数，λ_1，λ_2 は各損失項の重み係数である．この損失を導入することで，物理的・視覚的に自然な対応構造を自己教師型で学習可能とする．

4. 結果

図 1，図 2 ともに，緑を物体として，青を影として検出した領域とする．

図 1 は，提案手法による処理結果の成功例を示している．本例では，背景が比較的一様であり，物体の輪郭が明瞭であること，物体の色が明るく背景とのコントラストが十分であることから，成功したと考えられる．

一方，図 2 は失敗例を示しており，ステップ 2 において服や靴が影として誤検出されている．これは，初期の分離において明度・彩度の低さを影らしさの主要な指標として用いていることと，これらの領域が実際の影と空間的に近接していることが，領域の分離に失敗した理由であると考えられる．

5. まとめ

既存手法では，対応関係の学習には多量の教師データが必要とされる一方で，教師なしで動作する手法では意味的な区別や関係性の認識が困難である．

そこで，本研究では，AR における影生成の前段階として，影と物体の識別および対応関係の推定を，自己教師型学習の枠組みで行う手法を提案した．まず再構成誤差に基づいて特徴的な領域を検出し，その領域を色空間情報によって影と物体に分離する．さらに，空間的な関係に基づいて対応関係を推定することで，教師データを用いずに影と物体の意味づけと関係認識の実現を目指した．

今後は，検出精度や対応関係推定の妥当性について，影を含むシーンで構成される評価用のデータセットを用意し，影と物体の検出精度および対応関係推定の正確性を評価する．

参考文献

[1] Junsheng Xue et al. Lrgan: Learnable weighted recurrent generative adversarial network for end-to-end shadow generation. *Int. Joint Conf. Neural Netw. (IJCNN)*, 2024.

[2] Yan Hong et al. Shadow generation for composite image in real-world scenes. *Proc. AAAI Conf. Artif. Intell. 36*, 2022.

[3] Andres Sanin et al. Shadow detection: A survey and comparative evaluation of recent methods. *Pattern Recogn. 45(4)*, 2012.

[4] Tianyu Wang et al. Instance shadow detection with a single-stage detector. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3259–3273, 2023.

[5] Weisi Lin et al. Shadow detection using a cross-attentional dual-decoder network with self-supervised image reconstruction features. *Pattern Recogn. Lett.*, 175:1–8, 2024.

[6] Bruno Sauvalle et al. Autoencoder-based background reconstruction and foreground segmentation with background noise estimation. *arXiv:2112.08001*, 2021.