一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

マルチモーダル潜在空間を通じた 少数視点合成のための予測的意味正規化

松浦　史明†　　中溝　雄斗†　　田邊　　光†　　柳井　啓司†

† 電気通信大学　情報理工学域　I 類　〒182–8585 東京都調布市調布ケ丘一丁目 5 番地 1
E-mail: †{matsuura-f,nakamizo-y,tanabe-h}@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　少数画像からの新規視点合成は，限られた方向から撮影された画像のみを用いた場合，意味的な一貫性のみに基づいて学習すると，視点依存のアーティファクトが発生することが多い．本研究では，マルチモーダル大規模言語モデルである BLIP3-o を活用した，新たな NeRF フレームワークを提案する．未観測視点における粗い CLIP 特徴量を取得し，それらを NeRF の学習に組み込む．具体的には，(1) 入力各画像を CLIP エンコーダで特徴化し，(2) カメラ姿勢などの条件を含むテキストプロンプトにより BLIP3-o に目標未観測視点の意味特徴を予測させ，(3) 予測特徴と NeRF がレンダリングした画像の CLIP 特徴の類似度を損失として最適化する．さらに DietNeRF 由来のレンダリング損失および意味一貫性損失を併用し，外部深度・拡散モデル・手作業の正則化に依存せずに未観測表面のもっともらしい外観の再構成を可能とする．各種データセットでの定性的評価により，提案手法は背面アーティファクトを効果的に抑制した再構成を達成した．さらに定量評価では，構造的・知覚的整合性の向上を確認した．また，アブレーション実験により，色の忠実性や学習の安定性に関する課題も明らかとなった．

キーワード　3 次元再構成, Neural Radiance Fields, マルチモーダル大規模言語モデル

1. はじめに

1.1 背景・目的

3 次元再構成は，仮想現実（Virtual Reality, VR）や拡張現実（Augmented Reality, AR）分野において重要な技術である．特に，少数の視点画像から未知の視点を高精度で推定する技術は，没入感のある仮想体験やリアルな視覚表現を実現するために不可欠である．現在，この分野では Neural Radiance Fields（NeRF）[1] や 3D Gaussian Splatting (3DGS) [2] が注目を集めている．NeRF はニューラルネットワークを用いて 3 次元空間内の各点の色と密度を推定し，ボリュームレンダリングを通じて任意の視点からの画像を生成する技術である．しかし，NeRF を用いて高精度な視点再構成を行うには，多数の視点画像が必要であり，実用性の観点から課題が残されている．

特に，前方からの画像情報のみをもとに学習を行うと，背面にも同じような顔や目などの特徴が誤って生成されるという課題が存在する．そのため，従来の多くの手法は，このような問題を防ぐために，シーン内の視覚的一貫性を保つためには深度推定や各種の正則化が必要不可欠とされてきた．

他方，日常生活において，人間は視覚情報を意味に基づいて理解し，3 次元の深度を明示的に推定することなく自然にシーンを把握している．このことから，本研究では「物体の正面からの画像のみという極めて厳しい入力条件の元でも，原理的には，深度推定や複雑な正則化を用いず，画像の意味的理解のみを利用して不可視視点を含む自由視点映像を生成可能である」という仮説を置く．

そこで本稿では，マルチモーダル大規模言語モデル (Multimodal Large Language Model, MLLM) の一種である BLIP3-o [3] を用いて不可視視点の意味特徴を推定し，NeRF の学習に利用する新たな手法を提案する．従来手法では少数かつ偏った視点からの入力に対し，特に背面のような未観測領域で多くのアーティファクトが生じる問題があった．本研究はこの課題を，1 のように，BLIP3-o が推定した意味特徴を利用して改善し，少数の物体正面画像の入力から，未観測領域も含め視覚的に尤もらしい表現となるような自由視点映像の再構成を目指す．

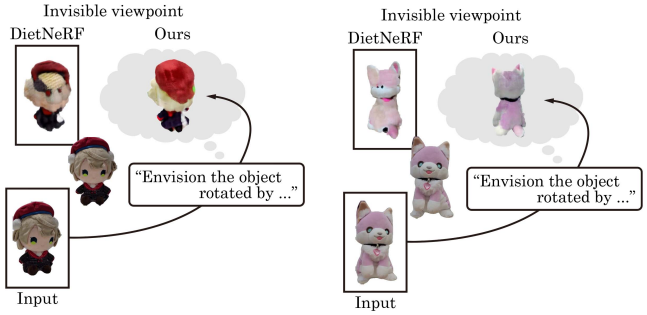


図 1: マルチモーダル大規模言語モデル (Multimodal Large Language Model, MLLM) の推論能力を活用して新規視点の意味的特徴を予測することにより，正面側画像のみを用いた未観測領域を含む 3 次元再構成が可能な手法を提案する．意味的特徴量を用いて学習の誘導を行っている従来手法では，背面側に本来存在し得ない目などが生成されていることが分かる．

本研究の主な貢献は以下の通りである．

- 少数ショット 3 次元再構成のための「視点条件付き意味

予測」を導入した．これは，単一の観測画像と姿勢を考慮したプロンプトから，MLLM を通して未観測視点の CLIP 特徴量を取得するものである．

- 予測されたそれぞれの視点の意味特徴を用いて NeRF の学習を誘導するための「意味予測損失（semantic prediction loss）」を提案し，プロンプト操作によっても未観測領域に対する言語主導の制御が可能であることを示した．

2. 関 連 研 究

2.1 少数ショットにおける 3 次元再構成

NeRF [1] や 3D Gaussian Splatting (3DGS) [2] は，画像のみからシーン表現を最適化できる一方で，一般物体に関する事前知識を内部に持たないため，少数ショット条件では未観測領域でアーティファクトが生じやすい．

2.2 正則化による最適化の安定化

少数ショットでのアーティファクトを抑える代表的アプローチは，NeRF の最適化に正則化を導入する方法である．意味的事前知識を利用する例として，レンダリング結果の意味整合を特徴空間で課す DietNeRF [4] や，幾何・色の正規化を併用する RegNeR [5] がある．幾何的事前知識を利用する例として，NeRF の推定深度から擬似教師を作る GeCoNeRF [6] や，複数スケールの整合性を深度条件で強める FrugalNeRF [7] がある．また，周波数成分の学習順序を制御して過学習を抑える FreeNeRF [8] も提案されている．ただし，これらは観測不足そのものを埋めるのではなく，あくまで最適化を破綻しにくくする試みであり，背面など未観測領域の外観を明示的に決める情報は不足しやすい．

2.3 未観測ビューの補完を介した誘導

別の潮流として，学習済み生成モデル等を用いて未観測領域を画像または潜在空間レベルで補完し，それを学習信号として NeRF に注入する方法がある．Text2NeRF [9] は，不可視視点の推定をテキスト条件付きインペイントとして扱い，拡散モデルで生成した画像と推定深度を学習に利用する．単一画像からの視点合成に基づく枠組みとして，Zero-1-to-3 [10] は視点条件付き拡散モデルの出力を用い，Score Jacobian Chaining (SJC) [11] により NeRF 空間を最適化することでゼロショット再構成を可能にした．これを基礎として，整合性強化を狙う Consistent123 [12] や段階的生成を行う Cascade-Zero123 [13] も提案されている．さらに ID-NeRF [14] は CNN で未観測領域を含む潜在特徴を推定し，拡散モデルによる蒸留を経て NeRF 学習をガイドする．しかし，生成による補完は未観測領域が尤もらしくなりやすい反面，視点間整合や観測画像への忠実度の保証が難しく，視点依存の破綻が残り得る．

2.4 ヤヌス問題

視点を変えると不自然な要素（例：背面に目が出る等）が現れる視点依存アーティファクトは，ヤヌス問題 (Janus problem) として指摘されている [15]．特に CLIP 特徴など観測画像由来の意味特徴で学習を誘導する場合，CLIP がカメラ姿勢や物体方位を明示的に保持しないこと，また回転や視点変換により埋め込みが大きく変化し得ることが報告されており [16]，特定視点に基づいた過学習が未観測視点での破綻を招きやすい．

2.5 本研究の位置づけ

本研究は，拡散モデルにより擬似ビューを生成して学習信号を補う Text-to-3D 系の枠組み [15], [17] とは異なり，見えている領域の外観・テクスチャを忠実に再構成したまま，未観測視点の情報不足を補うことを目指す．具体的には，観測画像と視点条件から MLLM により視点条件付きの意味特徴を抽出し，それを NeRF の正則化として導入することで，背面など未観測領域のアーティファクト抑制と少数ショット条件下での外観一貫性の向上を狙う．

3. 提 案 手 法

3.1 概　　要

前節までの課題に基づき，NeRF [1] および DietNeRF [4] を拡張した 3 次元再構成フレームワークを提案する．主要な新規点は，単一の入力画像と姿勢を考慮したプロンプトから BLIP3-o によって予測される，視点条件付きの意味特徴による NeRF 学習誘導の導入である．ランダムな姿勢からのレンダリング画像を，観測された学習画像の埋め込みに近づける DietNeRF とは異なり，本手法はレンダリングされる姿勢ごとに，ターゲット視点に依存した意味的な特徴量を生成する．この特徴量を学習の誘導して用いることで，観測された意味情報が未観測領域へ影響してしまう現象を緩和することができる．提案手法の概要を図 2 に示す．入力は，標準的な NeRF の設定に従い，複数の画像と，COLMAP [18] によって推定された擬似的なカメラポーズから構成される．

3.2 損 失 関 数

損失関数として，標準的なレンダリング損失と DietNeRF の意味的一貫性損失に加え，新たに**意味予測損失**（*semantic prediction loss*）を導入する．この損失は，ランダムな姿勢からの各レンダリング画像を，BLIP3-o がその特定の姿勢に対して予測した視点条件付き意味特徴へ整合させる方向へ学習を誘導させるものである．

3.2.1 レンダリング損失

NeRF [1] では，観測された画像とレンダリングされた画像の間のピクセル単位での整合性を保つための損失項が最適化の役割を果たしている．本手法でも同様に，サンプリングされた光線バッチ $\mathcal{R}$ に対して，ボリュームレンダリングによって得られた予測色 $\hat{C}(\mathbf{r})$ と，正解画像の色 $C(\mathbf{r})$ との間の平均二乗誤差（MSE）を最小化する．レンダリング損失 $\mathcal{L}_{\text{MSE}}$ は以下の式で定義される．

$$\mathcal{L}_{\text{MSE}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left\| C(\mathbf{r}) - \hat{C}(\mathbf{r}) \right\|_2^2 . \tag{1}$$

ここで，$\mathcal{R}$ は学習イテレーションごとにランダムにサンプリングされたレイの集合を表す．この損失により，観測済みの視点においては，高精細な形状とテクスチャの再構成が保証される．

3.2.2 意味一貫性損失

DietNeRF [4] で提案された意味一貫性損失は，事前学習済みの画像エンコーダを用いて，任意の視点からのレンダリング画

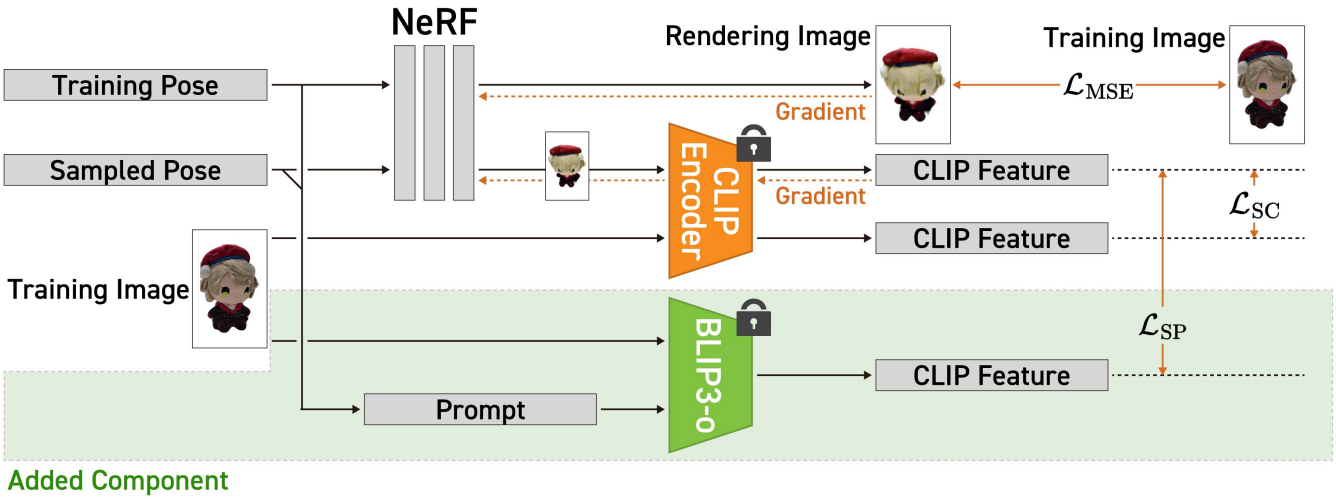


図 2: 提案手法の概要

像が，学習画像と意味的に類似することを強制する正則化項である．具体的には，学習データに含まれる画像 I と，ランダムにサンプリングされた未観測の視点 $\mathbf{p}_{\text{r}}$ からレンダリングされた画像 $\hat{I}_{\mathbf{p}_{\text{r}}}$ をそれぞれエンコーダ ϕ に入力し，その特徴空間上での距離を最小化する．意味一貫性損失 $\mathcal{L}_{\text{SC}}$ は以下の式で表される．

$$\mathcal{L}_{\text{SC},\ell_2}\left(I, \hat{I}_{\mathbf{p}_{\text{r}}}\right) = \left\|\phi\left(I\right) - \phi\left(\hat{I}_{\mathbf{p}_{\text{r}}}\right)\right\|_2^2. \tag{2}$$

この損失項に対し，ハイパーパラメータ λ_{sc} を乗じたものが最終的な損失関数に加算される．これにより，ピクセルレベルの対応が取れない未観測視点においても，物体の意味的な整合性が保たれるよう学習が誘導される．DietNeRF では OpenAI の CLIP を画像エンコーダとして用いていたが，本研究では EVA-CLIP [19] を用いた．これは，BLIP3-o で中間表現として生成される CLIP が EVA-CLIP 準拠のものであったことから，類似度を算出し比較可能とするためである．

3.2.3 意味予測損失

新たに，意味予測損失を導入する．ランダムに決定されたポーズ $\mathbf{p}_{\text{r}}$ からのレンダリング画像 $\hat{I}_{\mathbf{p}_{\text{r}}}$ の EVA-CLIP 特徴 $\phi(\hat{I}_{\mathbf{p}_{\text{r}}})$ と，訓練画像 I を用いてポーズ $\mathbf{p}_{\text{r}}$ からの視点でどのように見えるかを BLIP3-o(B) を用いて予測した CLIP 特徴 $B(I, \text{prompt})$ の類似度を最小化するような損失を定義する．以降，本損失を Semantic prediction loss とし，次の式で与えられるものとする．

$$\mathcal{L}_{\text{SP},\ell_2}\left(I, \hat{I}_{\mathbf{p}_{\text{r}}}\right) = \left\|B(I, \text{prompt}) - \phi\left(\hat{I}_{\mathbf{p}_{\text{r}}}\right)\right\|_2^2. \tag{3}$$

この損失項に対し，ハイパーパラメータ λ_{sp} を乗じたものが最終的な損失関数に加算される．なお，本損失にあたり，$B(I, \text{prompt})$ は画像へのデコードは行わない．BLIP3-o [3] のようにデコードを行うには，追加の CLIP 条件付き拡散デコーダーが必要となり，拡散モデルの生成品質への依存と計算量の増大が生じるためである．その代わり，$B(I, \text{prompt})$ は特徴レベルの監督信号としてのみ使用される．勾配は $\phi(\hat{I}_{\mathbf{p}_{\text{r}}})$ を通じて NeRF レンダラーへと流れるため，予測された特徴は再構成されるテクスチャに直接影響を与える．実際には，色やテクスチャの忠実度を損なわないよう，$\mathcal{L}_{\text{SP}}$ は適度な重み λ_{sp} を持つ補助的な正則化項として扱う．

MLLM へのプロンプト本手法では，BLIP3-o が「視点が変化した際にどのように物体が見えるか」という未知の視点に対する意味特徴を予測できるようにするため，学習画像とターゲットレンダリング姿勢との間の相対的なカメラ回転を自然言語プロンプトへエンコードする．具体的には，訓練画像とレンダリング対象視点との相対回転角（Yaw, Pitch, Roll）を抽出し，それを自然言語で記述したプロンプトを生成する．これにより，単なる説明文ではなく，特定の視点変化に対応する「意味的条件付き入力」として BLIP3-o に提示され，視点変化を反映した CLIP 特徴の推定を可能としている．

不可視視点の CLIP 特徴予測値 $B(I, \text{prompt})$ を求めるためのプロンプトとして，回転角度を自由に組み込むことができる，図 3 に示すテキストを用意した．テンプレートは英語だが，指定された視点から見た情景を想像し，画像を生成することを意図したプロンプトである．

> *Please generate image based on the following caption:*
> *Envision the central object rotated by {yaw:.0f}° to the right, {pitch:.0f}° upward, and {roll:.0f}° clockwise; describe its overall appearance—including shape, color, and texture—and specify any objects or features that would be absent, hidden, or not included from this viewpoint.*

図 3: 不可視視点予測のためのプロンプト

プロンプトを MLLM へ入力するにあたりまず，訓練画像 I のポーズ $\mathbf{p}$ の回転行列 $R_{\mathbf{p}}$ と，レンダリング画像 $\hat{I}_{\mathbf{p}_{\text{r}}}$ のポーズ $\mathbf{p}_{\text{r}}$ の回転行列 R_{target} から相対回転行列 $R_{\text{rel}} = R_{\mathbf{p}} {R_{\text{target}}}^T$ を計算することで，訓練画像 I と $\hat{I}_{\mathbf{p}_{\text{r}}}$ の視点の角度差を求めた．これにより得られた Yaw（右回転方向），Pitch（上向き方向），Roll（時計回り方向）の角度情報を図 3 に示すテンプレートに埋め込む．このプロンプトを学習画像 I と組み合わせることで，BLIP3-o は新しい視点からの物体の外観を反映した条件付き予測 CLIP 特徴 $B(I, \text{prompt})$ を出力することが可能となる．

3.2.4 全体の損失関数

以上の損失を統合し，本手法の損失関数は次式で示すものと

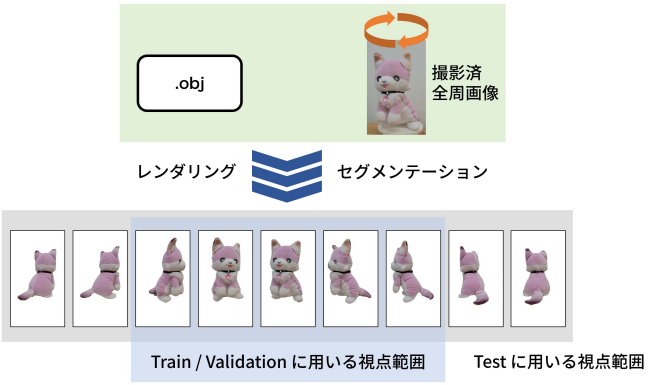


図 4: データセット作成の概要

した．

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{sc}}\mathcal{L}_{\text{SC},\ell_2} + \lambda_{\text{sp}}\mathcal{L}_{\text{SP},\ell_2}. \quad (4)$$

4. 実　　験

4.1 実験設定

4.1.1 データセット

本実験では，独自に作成したデータセットと，Google Scanned Objects (GSO) データセット [20] の 2 種類を使用した．

データセット作成時の条件 NeRF の学習には画像とそれに対応するカメラポーズが必要である．図 4 に示すように，実験対象のデータとして前面-背面，右側面-左側面で形が対称でない 3D オブジェクトをデータセット作成対象として選定し，それを用いて物体を 360 度全周から撮影，あるいはレンダリングした画像を再構成モデル訓練のためのデータセットとした．データセットは学習の際に用いる train データ，学習時にパラメータ調整アルゴリズムにより用いられる validation データ，学習のモデルを評価するために用いる test データの 3 種類で構成されているが，train と validation に関しては全周画像をすべて用いるのではなく，物体の正面と考えられる位置からカメラのヨー角が約 180° ± 90° ずれた視点を除いたものを用いた．これら「正面ビュー」のみを用いた上で，背面の整合性を取りながら再構築可能かどうかを test データを用いて評価を行った．

4.1.2 評価指標

評価指標は，画素誤差に基づく PSNR，局所構造の類似度を測る SSIM [21]，深層特徴に基づく知覚距離 LPIPS [22] を用いて，いずれも学習に用いていない test 画像に対して算出した．

4.1.3 実験パラメータの設定

実験は通常の NeRF，DietNeRF，DietNeRF (eva-clip)（CLIP を EVA-CLIP [19] に置き換え，他の設定は同一とした DietNeRF），および提案手法である 4 通りのパターンで，表 1 に示す条件で行った．本研究の主な焦点は意味情報のみを用いた学習の有効性を検証することにあるため，幾何的や情報論的などその他の正則化を含まず，意味ベースの学習誘導の効果を直接評価している DietNeRF をベースラインとして選択した．表 1 に示す条件以外，$\lambda_{\text{sc}} = 0.1$ を含むすべてのハイパーパラメータは DietNeRF に合わせて設定した．学習イテレーションは妥当な表現を得るために最低限必要である 20,000 回とし，評価も 20,000 回時点でのモデルを用いて行った．なお，表中の #Train は学習用画像から実際に学習に用いる画像の枚数，$Freq_{\text{sc}}, Freq_{\text{sp}}$ はそれぞれ，学習イテレーションでの意味一貫性損失と意味予測損失の適用間隔である．

表 1: 実験設定

Method	#Train	Optimizer	Image Encoder ϕ	λ_{sc}	$Freq_{\text{sc}}$	λ_{sp}	$Freq_{\text{sp}}$
NeRF	8	Adam	-	-	-	-	-
DietNeRF	8	Adam	CLIP-ViT-B-32	0.1	10	-	-
DietNeRF (eva-clip)	8	Adam	EVA-CLIP-E-14-plus	0.1	10	-	-
Ours	8	Adam	EVA-CLIP-E-14-plus	0.1	10	0.05	10

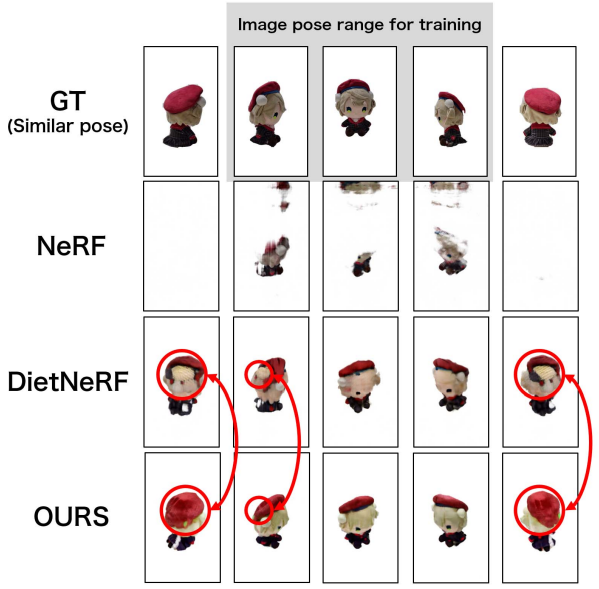


図 5: 提案手法を用いた生成結果の例

4.2 実　　験

4.2.1 定性評価

図 5 は，独自データセットに本手法を適用した結果の例である．DietNeRF ではぬいぐるみの背面に偽の目や口などのアーティファクトが生じているのに対し，提案手法ではそれが発生していないことが分かる．一方で，BLIP3-o で予測された CLIP 特徴には色の保持に課題がある（詳細は後述）ことから，アーティファクトを消すことができているものの，周囲の色とは異なる偽色が発生することがある．

4.2.2 定量評価

それぞれの手法を用いて，背面を含まない学習用データから，背面を含むテスト用データで各種評価指標を計測した．学習を行ったすべてのデータ（独自データと GSO データを合わせた，計 28 種類のデータ）に対しての平均値は表 2 の通りとなった．

LPIPS, SSIM の値が一貫して他手法を上回ることが確認された．一方で PSNR について，他手法が優れていることがあることが分かった．詳しい分析は次節以降へ記述する．

4.2.3 まとめ

以上の定性・定量評価から，提案手法の有効性として次の点が分かった．(1) 視点依存アーティファクトの抑制：DietNeRF で見られた背面の顔の出現がほぼ解消されたこと．(2) 構造的・

表 2: 各種評価指標の平均値とその比較（すべてのデータセットの平均）

Method	PSNR↑	SSIM↑	LPIPS↓
NeRF	24.9	0.925	0.100
DietNeRF	**25.6**	0.924	0.094
DietNeRF (eva-clip)	23.5	0.922	0.098
Ours	24.2	**0.927**	**0.092**

知覚的整合性の向上：全データセット平均で SSIM，LPIPS で最良値を達成し，未観測領域の学習誘導が適切に行われていること．(3) 色再現性とのトレードオフ：PSNR は DietNeRF や NeRF に劣るが，CLIP 特徴量は意味情報を重視するため，画素値の完全一致よりも不自然なアーティファクトの解消と尤もらしい 3 次元形状の再構成を優先していること．総じて，提案手法は未観測領域におけるアーティファクト抑制と知覚的に妥当な形状復元の点で既存手法を上回ることが確認された．

4.3 アブレーション分析

4.3.1 意味予測損失のみを用いた場合

図 6 の (c) は，意味予測損失のみを適用し，3 次元再構成を試みた結果である．NeRF に比べ，不可視視点である背面の形も含めて一貫して推測できていることが分かるが，色やテクスチャを保持できていないことが分かる．この結果より，BLIP3-o により予測された CLIP 特徴は，視点変化による物体の形状変化については推測できているものの，細かな色やテクスチャについては推測できていないことが分かる．

4.3.2 MLLM へ与えるプロンプトを変更したときの変化

前項までは，尤もらしい不可視視点を想像させるため，プロンプトには学習画像と視点推移の角度情報を組み合わせていた．このプロンプトを変更することで，レンダリング結果にどのような変化が生じるかを検証した．今回は，プロンプトを「パンダの画像」と変更した．レンダリング結果は図 6 の (d) の通りとなった．正面は学習に用いた画像の通りとなっているが，不可視視点で想像の必要がある背面は，プロンプト通りパンダのような白黒のテクスチャが現れる結果となった．

以上の結果より，λ_{sp} は NeRF 学習に対して単なるノイズを引き起こしていたわけではなく，正しく BLIP3-o による予測結果に基づいて，学習を誘導できていることが確認された．また，BLIP3-o の推測による影響が，未観測領域の NeRF 再構成結果にのみ影響を与えていることが分かる．

4.3.3 意味予測損失の重みを変化させたときの変化

意味予測損失の，損失関数全体に与える影響を調査するため，λ_{sp} を 0.0 から 1.0 まで変化させた場合の，各種評価指標の推移を図 7 に示す．なお，λ_{sc} については実験全体で 0.1 に固定をしている．独自データセット全体の PSNR (図 7a)，SSIM (図 7b)，LPIPS (図 7c) の平均値の推移は図の通りとなった．

結果より，λ_{sp} が 0.05 の場合に，SSIM と LPIPS が向上することが分かる．一方で，PSNR は λ_{sp} が 0.15 のときに最も向上する結果となった．意味予測損失のみを用いた場合の結果と合わせて考えると，可視視点のレンダリング結果を損なうことなく不可視視点を補完するためには，ある程度 λ_{sc} を用いて学習が進んでいるところに，補助的に損失を挿入することが必要であることが分かる．

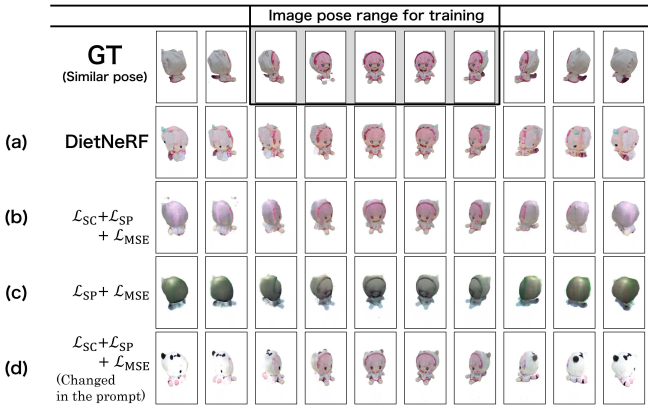
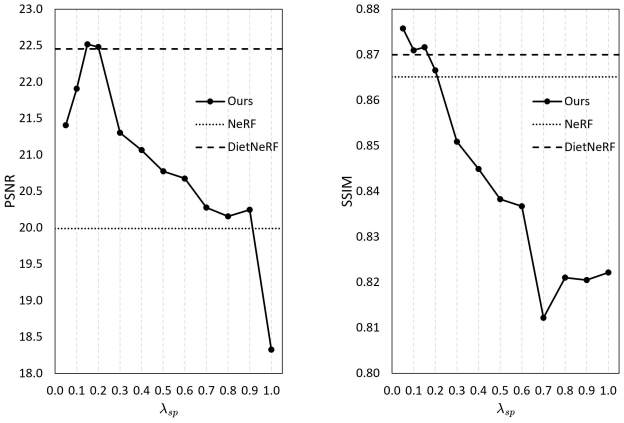
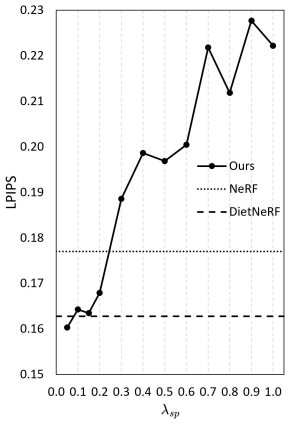


図 6: アブレーション分析．(a) DietNeRF，(b) 提案手法，(c) 意味予測損失のみを用いた場合，(d) 提案手法でプロンプトを「パンダ」に変えた場合．



(a) λ_{sp} の変化による PSNR の推移　(b) λ_{sp} の変化による SSIM の推移



(c) λ_{sp} の変化による LPIPS の推移

図 7: λ_{sp} の変化に対する各評価指標の平均値の推移

5. 考　　察

5.1 言語空間による構造理解の向上

従来の拡散モデルベースの手法では，視点補完や欠損領域の補間をピクセル空間または特徴空間上で行うことが一般的であったが，本手法では自然言語プロンプトを介した視点条件付きの意味的な学習誘導を導入する．この条件付けにより，観測視点のバイアスに起因する背面アーティファクトの発生を抑制しやすくなる．

5.2 PSNR の低下

表 2 より，提案手法は知覚的・構造的指標では比較手法を上回った一方で，PSNR はベースラインである DietNeRF より低い．また定性的にも，図 7 に示されるように，$\mathcal{L}_{\mathrm{SP}}$ に過度に依存すると色の整合性が損なわれる．これは，少なくとも本実験の条件においては，BLIP3-o が生成する CLIP 特徴量が細かな色再現性を重視していないことを示唆している．したがって，色とテクスチャの一貫性を維持することを目指す場合，$\mathcal{L}_{\mathrm{SP}}$ は主要な目的関数としてではなく，適度な補助信号（すなわち，小さな λ_{sp} を用いる）として扱われるべきである．ピクセルレベルの完全な再構成よりも高レベルな意味的な整合性の確保を優先したことが，DietNeRF と比較して PSNR がわずかに低下した要因である．しかし，本研究の主目的は正確なピクセルレベルでの再構成品質の向上を達成することではなく，未観測領域におけるアーティファクトを解消することにあるため，このトレードオフは許容範囲内であると考える．SSIM および LPIPS スコアの向上は，PSNR が低いにもかかわらず，結果が知覚的により尤もらしいことを裏付けている．

5.3 計算コスト

MLLM の組み込みは推論のオーバーヘッドをもたらすが，提案手法の設計は計算コスト低減を優先し，BLIP3-o の画像生成プロセスは中間 CLIP 特徴の段階で停止することとした．完全な画像生成に必要な計算コストの高い拡散デコーダを経由しないことで，完全な擬似視点を合成する手法と比較して，追加コストを低く抑えることができている．

6. ま と め

本研究では，少数視点入力時に発生する背面アーティファクトに対し，MLLM を用いて視点条件付きの意味特徴を予測し，NeRF 学習を誘導する手法を提案した．定性評価によりアーティファクトの抑制を確認し，定量評価でも全データ平均で構造および知覚品質の改善を確認した．

一方で，色再現性・高周波テクスチャの保持は難しく，意味誘導の有効性とピクセル単位での忠実度の両立には課題が残る．今後は，意味特徴に加え，画像生成モデルを補助的に併用する，あるいはテクスチャをより反映する正則化設計・高解像度特徴を持つ MLLM の導入により，構造的正しさと外観忠実度の両立を目指す．

文　　献

[1] B. Mildenhall, and P.P. Srinivasan, et al., “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis,” Proc. of European Conference on Computer Vision, pp.405–421, 2020.

[2] B. Kerbl, and G. Kopanas, et al., “3D Gaussian Splatting for Real-Time Radiance Field Rendering,” ACM Transactions on Graphics, vol.42, no.4, pp.139:1–139:14, 2023.

[3] J. Chen, and Z. Xu, et al., “BLIP3-o: A Family of Fully Open Unified Multimodal Models-Architecture, Training and Dataset,” arXiv preprint arXiv:2505.09568, 2025. https://arxiv.org/abs/2505.09568

[4] A. Jain, and M. Tancik, et al., “Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis,” Proc. of IEEE International Conference on Computer Vision, pp.5865–5874, 2021.

[5] M. Niemeyer, and J.T. Barron, et al., “RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs,” Proc. of IEEE Computer Vision and Pattern Recognition, pp.5480–5490, 2022.

[6] M. Kwak, and J. Song, et al., “GeCoNeRF: Few-shot neural radiance fields via geometric consistency,” Proc. of International Conference on Machine Learning, vol.202, pp.18023–18036, Proceedings of Machine Learning Research, 2023.

[7] C. Lin, and C. Wu, et al., “FrugalNeRF: Fast convergence for extreme few-shot novel view synthesis without learned priors,” Proc. of IEEE Computer Vision and Pattern Recognition, pp.11227–11238, 2025.

[8] J. Yang, and M. Pavone, et al., “FreeNeRF: Improving Few-shot Neural Rendering with Free Frequency Regularization,” Proc. of IEEE Computer Vision and Pattern Recognition, pp.8254–8263, 2023.

[9] J. Zhang, and X. Li, et al., “Text2NeRF: Text-driven 3d scene generation with neural radiance fields,” IEEE Transactions on Visualization and Computer Graphics, vol.30, no.12, pp.7749–7762, 2024.

[10] R. Liu, and R. Wu, et al., “Zero-1-to-3: Zero-shot one image to 3d object,” Proc. of IEEE International Conference on Computer Vision, pp.9264–9275, 2023.

[11] H. Wang, and X. Du, et al., “Score Jacobian Chaining: Lifting pretrained 2d diffusion models for 3d generation,” Proc. of IEEE Computer Vision and Pattern Recognition, pp.12619–12629, 2023.

[12] H. Weng, and T. Yang, et al., “Consistent123: Improve consistency for one image to 3d object synthesis,” arXiv preprint arXiv:2310.08092, 2023.

[13] Y. Chen, and J. Fang, et al., “Cascade-Zero123: One image to highly consistent 3d with self-prompted nearby views,” Proc. of European Conference on Computer Vision, vol.15099, pp.311–330, Lecture Notes in Computer Science, 2024.

[14] Y. Li, and S. Wang, et al., “ID-NeRF: Indirect diffusion-guided neural radiance fields for generalizable view synthesis,” Expert Systems with Applications, vol.266, p.126068, 2025.

[15] B. Poole, and A. Jain, et al., “DreamFusion: Text-to-3d using 2d diffusion,” Proc. of International Conference on Learning Representations, 2023.

[16] A. Dahal, and S.A. Murad, et al., “Embedding shift dissection on CLIP: Effects of augmentations on vlm’s representation learning,” Proc. of IEEE Computer Vision and Pattern Recognition Workshops, pp.4853–4857, 2025.

[17] C. Lin, and J. Gao, et al., “Magic3D: High-resolution text-to-3d content creation,” Proc. of IEEE Computer Vision and Pattern Recognition, pp.300–309, 2023.

[18] J.L. Schönberger and J. Frahm, “Structure-from-Motion Revisited,” Proc. of IEEE Computer Vision and Pattern Recognition, pp.4104–4113, 2016.

[19] Q. Sun, and Y. Fang, et al., “EVA-CLIP: Improved Training Techniques for CLIP at Scale,” arXiv preprint arXiv:2303.15389, 2023.

[20] L. Downs, and A. Francis, et al., “Google scanned objects: A high-quality dataset of 3d scanned household items,” Proc. of IEEE International Conference on Robotics and Automation, pp.2553–2560, 2022.

[21] Z. Wang, and A.C. Bovik, et al., “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, vol.13, no.4, pp.600–612, 2004.

[22] R. Zhang, and P. Isola, et al., “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric,” Proc. of IEEE Computer Vision and Pattern Recognition, pp.586–595, 2018.