# BlipNeRF: MLLMによる不可視視点の予測を用いた少数の物体正面画像からの3次元再構成

松浦 史明[1,a)]　中溝 雄斗[1,b)]　田邊 光[1,c)]　柳井 啓司[1,d)]

## 概要

本研究では，マルチモーダル大規模言語モデル（MLLM）の一種である BLIP3-o[1] を用いて不可視視点を推定し，少数の物体正面画像のみから自由視点映像を再構成する手法を提案する．実験により，既存手法の特定の条件における課題であった背面アーティファクトの抑制と視覚的一貫性の向上を確認した．

## 1. はじめに

3 次元再構成は，仮想現実（VR）や拡張現実（AR) 分野において重要な技術である．特に，入力画像からは見ることのできない視点が存在する場合，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．しかし，これらの手法は計算量と設計難易度の面で高いコストを伴う．

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，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，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考えられる．

そこで本稿では，MLLM の一種である BLIP3-o 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である BlipNeRF 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．本研究の BlipNeRF はこの課題を，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．

[1] 電気通信大学
[a)] matsuura-f@mm.inf.uec.ac.jp
[b)] nakamizo-y@mm.inf.uec.ac.jp
[c)] tanabe-h@mm.inf.uec.ac.jp
[d)] yanai@cs.uec.ac.jp

## 2. 関連研究

任意の視点から画像を生成することを目的とした新規視点合成の手法として NeRF [5] や 3D Gaussian Splatting (3DGS) [4] などの手法が存在する．どちらも手法も学習時に用いるのはシーンに関する画像のみのため，一般的な物体に対する事前知識を活用することができない．そのため，高品質な視点合成を行うためには多数のシーンの観測データを必要とするが，入力画像が少数の場合，新規視点では多数のアーティファクトが生じる傾向がある．

これらの問題を解決するため，様々な正則化手法が提案されている．DietNeRF [2] では，レンダリング結果の CLIP [7] 特徴空間での意味的類似性を高める手法を，RegNeRF [6] では深度平滑化や色正則化を，FreeNeRF [12] では学習時の入力画像の周波数範囲の正則化を，MutualNeRF [11] では視点間の相互情報量に注目することで，少数視点映像からの高品質な３次元再構成を目指している．

### 2.1 DietNeRF

少数視点入力のレンダリング品質向上のために提案された DietNeRF ではあるが，入力画像が偏った視点のみの画像である場合，未観測領域では多数のアーティファクトが生じるという課題がある．特に．正面と背面でテクスチャ等の配置が対称でない場合に発生しやすい．

#### 2.1.1 レンダリング損失

NeRF では，レンダリング損失としてピクセル単位での MSE（平均二乗誤差）を最小化するよう，以下の手順で学習が行われる．

(1) 訓練画像とポーズのペア $(I, \mathbf{p}_i)$ をランダムにサンプリングする．

(2) 同じポーズ $\mathbf{p}_i$ からのレンダリング画像 $\hat{I}_{\mathbf{p}_i}$ をボリュームレンダリング [3] によって生成する．このとき，各ピクセルの色は [2] の式 1 により次の式で表される．

$$\hat{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\,\sigma\big(\mathbf{r}(t)\big)\,c\big(\mathbf{r}(t), d\big)\,\mathbf{d}t \qquad (1)$$

このとき，$T(t)$ は光線が $t_n$ から $t$ まで遮られずに進

む確率であり，次式で表される．

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))\, \mathrm{d}s\right) \qquad (2)$$

また，$\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ は，カメラ位置 $o$ から方向ベクトル $d$ に沿った 3 次元の線分（レイ），$\sigma(\mathbf{r}(t))$ は点 $\mathbf{r}(t)$ の不透明度，$c(\mathbf{r}(t)$ は点 $\mathbf{r}(t)$ の方向 $d$ に対する RGB 値，$t_n, t_f$ はレイの近点，遠点である．

(3) 一部のピクセル（レイ）をサンプリングし，[2] の式 3 に示される通り，MSE を最小化する．

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left\| C(\mathbf{r}) - \hat{C}(\mathbf{r}) \right\|_2^2 \qquad (3)$$

このとき，$\mathcal{R}$ はレイの全体集合，$C(\mathbf{r})$ と $\hat{C}(\mathbf{r})$ はそれぞれ色の真値と推定値である．

入力視点数が多い場合にはこの損失関数は有効であるが，視点数が少ない場合，とりわけ偏った視点の入力である場合は，入力視点近傍の結果のみを最適化しようとするあまり過学習に陥ることがある．このとき，不可視視点からの映像を再構成すると不透明な領域が視線を遮り，物体がレンダリングされない領域を発生させることになる．

#### 2.1.2 意味一貫性損失

前述のレンダリング損失に加え，DietNeRF では事前学習された画像エンコーダから得られる知識を活用するため，任意のカメラポーズにおける損失として，意味一貫性損失が導入された．

異なる視点から得られた画像でも，同じ物体であれば意味的に同じになるはずである，という考え方から，訓練画像 $I$ と，ランダムに決定されたポーズ $\mathbf{p}_{\mathrm{r}}$ からのレンダリング画像 $\hat{I}_{\mathbf{p}_{\mathrm{r}}}$ の意味的類似度を最小化する [2] の式 4 により示される以下の損失が提案された．このとき，$\phi$ は CLIP などの画像エンコーダ，$\lambda_{\mathrm{sc}}$ は損失の重みである．

$$\mathcal{L}_{\mathrm{SC},\ell_2}\left(I, \hat{I}_{\mathbf{p}_{\mathrm{r}}}\right) = \lambda_{\mathrm{sc}} \left|\left|\phi\left(I\right) - \phi\left(\hat{I}_{\mathbf{p}_{\mathrm{r}}}\right)\right|\right|_2^2 \qquad (4)$$

再構成を試みている物体にある程度の対称性が存在する場合にはこの損失関数は有効であるが，偏った視点の入力である場合，ある一方の視点に不可視視点のレンダリング結果が誘導されてしまう．そのため，正面には存在するが，背面には存在し得ないテクスチャ（人間の場合，目や口などは正面に存在するが，背面には存在しない）がアーティファクトとして出現しやすい．

### 2.2 BLIP3-o

他方，画像を条件づけとし，テキストから画像を生成する MLLM は様々あるが，生成過程の中間表現として CLIP を用いるものとして，BLIP3-o [1] が挙げられる．

BLIP3-o では，画像生成を行う際に，自己回帰モデルと

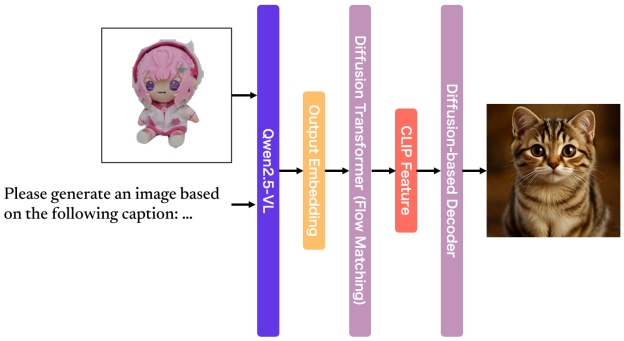


**図 1** BLIP3-o の画像生成フロー

拡散トランスフォーマを組み合わせた構造を採用している．具体的には，図 1 に示すように，入力→自己回帰モデルによる視覚特徴の生成→拡散モデルによる CLIP 特徴空間への変換→拡散モデルによる RGB 画像への変換，という複数の過程を踏んでいる．

これを利用することで，不可視視点を想像したときにどのような CLIP 特徴となるかをある程度予測させることができる．例えば，物体の正面の画像と「背面はどうなりますか？」というプロンプトを入力し，画像生成フローを辿ると，生成画像の元となる CLIP 特徴を取得することができる．

これにより，不可視視点の完全な画像生成が困難な MLLM でも，視点変化のおおまかな特徴が反映された CLIP 特徴を取得することができる．

## 3. 手法

これまでの課題を踏まえ，DietNeRF をベースとし，プロンプトを用いて視点変化を予測できる BLIP3-o を組み合わせた，新たな NeRF 再構成手法を提案する．

概要を図 2 に示す．入力は NeRF と同様に複数の画像と COLMAP で求められた画像の擬似的なカメラポーズである．

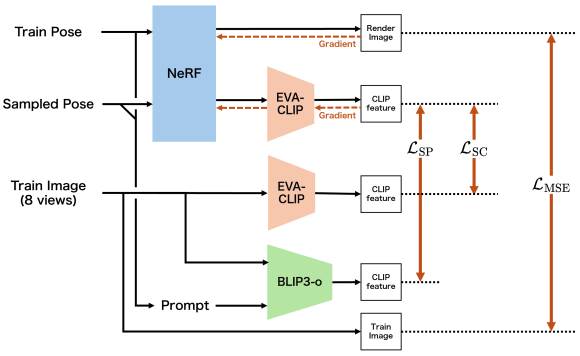


**図 2** 提案手法の概要

損失関数として，DietNeRF と同様にレンダリング損失と意味一貫性損失を導入するとともに，新たに意味予測損失を導入する．

### 3.1 意味一貫性損失

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 [9] を用いた．これは，既存の CLIP よりも EVA-CLIP の方が表現力が高いことと，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，類似度を算出しやすくするために用いた．以上の意味一貫性損失関数は，$\phi$ を EVA-CLIP としたこと以外，式 4 と同様である．

### 3.2 意味予測損失

次に，意味予測損失を導入する．ランダムに決定されたポーズ $\mathbf{p}_\mathrm{r}$ からのレンダリング画像 $\hat{I}_{\mathbf{p}_\mathrm{r}}$ の EVA-CLIP 特徴 $\phi(\hat{I}_{\mathbf{p}_\mathrm{r}})$ と，訓練画像 $I$ を用いてポーズ $\mathbf{p}_\mathrm{r}$ 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 $B(I, prompt)$ 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．なお，$\lambda_\mathrm{sp}$ は損失の重みである．

$$\mathcal{L}_{\mathrm{SP},\ell_2}\left(I, \hat{I}_{\mathbf{p}_\mathrm{r}}\right) = \lambda_\mathrm{sp}\ ||B(I, prompt) - \phi\left(\hat{I}_{\mathbf{p}_\mathrm{r}}\right)||_2^2 \quad (5)$$

#### 3.2.1 MLLM へのプロンプト

不可視視点の CLIP 特徴予測値 $B(I, prompt)$ 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次のテキストを用意した．

*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

訓練画像 $I$ のポーズ $p$ の回転行列 $R_\mathbf{p}$ と，レンダリング画像 $\hat{I}_{\mathbf{p}_\mathrm{r}}$ のポーズ $\mathbf{p}_\mathrm{r}$ の回転行列 $R_\mathrm{target}$ から相対回転行列 $R_\mathrm{rel} = R_\mathbf{p} {R_\mathrm{target}}^T$ を計算することで，訓練画像 $I$ と $\hat{I}_{p_r}$ 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）の角度情報をプロンプトとし，画像 I を BLIP3-o へ入力することで，$B(I, prompt)$ の生成を行った．

### 3.3 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した．

$$\mathcal{L}_\mathrm{total} = \mathcal{L}_\mathrm{MSE} + \mathcal{L}_\mathrm{SC} + \mathcal{L}_\mathrm{SP} \quad (6)$$

## 4. 実験

### 4.1 データセット

学習には，独自に作成した 12 種類のぬいぐるみの画像からなるデータセットを使用した．


**図 4** データセット作成の概要

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が，まず画像について，図 4 に示すように，ぬいぐるみの外観を全周から撮影し，YOLOv8 [10] を使用してぬいぐるみのみを切り抜いたものを使用し，ランダムに train データ，validation データ，test データとした．train データと validation データについては背面画像を除き，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．カメラポーズについて，Structure from Motion (SfM) 手法の一種である COLMAP [8] を用いて，擬似的なカメラポーズの推定を行った．

### 4.2 実験条件

実験は通常の NeRF，DietNeRF，提案手法の 3 通りで，表 1 に示す条件で行った．なお，interval はそれぞれの損失関数を適用する間隔である．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 20000 回とし，評価も 20000 イテレーション時点でのモデルを用いて行った．

**表 1** 実験条件

| Method | 学習画像 | Optimizer | 画像エンコーダ $\phi$ | $\lambda_{sc}$ | $\mathcal{L}_{SC}$ interval | $\lambda_{sp}$ | $\mathcal{L}_{SP}$ interval |
|---|---|---|---|---|---|---|---|
| NeRF | 8 枚 | Adam | - | - | - | - | - |
| DietNeRF | 8 枚 | Adam | clip-ViT-B-32 | 0.1 | 10 | - | - |
| Ours | 8 枚 | Adam | eva-clip-E-14-plus | 0.1 | 10 | 0.15 | 10 |

### 4.3 実験

#### 4.3.1 定性評価

図 5，図 6 は前出のデータセットに本手法を適用した例である．どちらの例も DietNeRF では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，図 6 のように周囲の色とは異なる偽色が発生することがある．

#### 4.3.2 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．12 種類中，学習に成功した 10 種類の平均値は表 2 の通りとなった．なお，残りの 2 種類については学習が安定せず，レンダリング結果が崩壊したため除外を行った．

#### 4.3.3 まとめ

以上の結果より，従来の意味一貫性では対応が困難であった背面方向の整合性について，本手法を通して違和感の少ない状態で再構成が可能であることが確認された．


**図 5**　提案手法を用いた生成結果の例 1

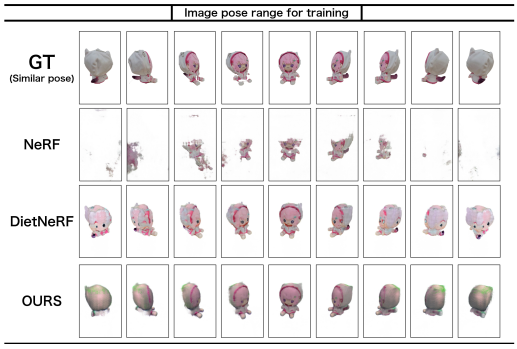


**図 6**　提案手法を用いた生成結果の例 2

**表 2**　各種評価指標の平均値とその比較

| Method | PSNR↑ | SSIM↓ | LPIPS↑ |
|---|---|---|---|
| NeRF | 19.9 | 0.865 | 0.177 |
| DietNeRF | 22.4 | 0.872 | **0.163** |
| Ours | **22.5** | **0.870** | 0.164 |

### 4.4　アブレーション分析

#### 4.4.1　BLIP3-o の想像力の確認

実験にあたり，BLIP3-o の不可視視点に対する予測能力を確認するため，図 7 に示す通り任意の学習画像 1 枚とプロンプトを入力し，その応答を確認した．

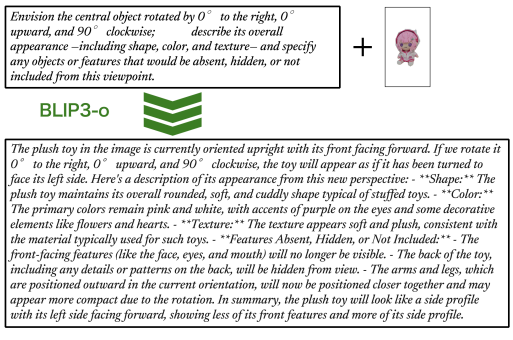


**図 7**　BLIP3-o による視点変化の想像力の確認

応答は極めて自然であり，視点変化を起こした場合のおおまかな物体の見え方を想像できていることが分かる．この結果より，画像生成の過程でも同様に，物体の見えを想像した結果をある程度反映した CLIP 特徴量を生成可能だと判断した．

#### 4.4.2　意味予測損失のみを用いた場合

図 8 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．

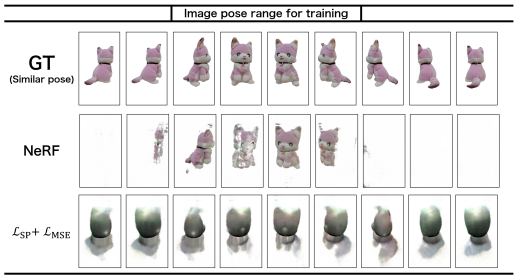


**図 8**　意味予測損失のみを用いた生成結果の例

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．

## 5.　議論

### 5.1　CLIP 予測値の妥当性

本手法は，BLIP3-o の CLIP 予測値に学習結果が大きく依存するものである．形状の推測については優れた性能を発揮している一方で，色やテクスチャの保持ができない課題については，プロンプトや使用する MLLM を変更することで改善できる可能性がある．

### 5.2　学習の安定性

本実験では 12 種類のぬいぐるみを撮影したデータセットを用いたが，そのうち 10 種類の学習に成功し，2 種類の学習に失敗する結果となった．このことより，$\mathcal{L}_{\mathrm{SP}}$ の適用が学習の安定性をある程度損ねていると考えることができる．

### 5.3　おわりに

本稿では，既存の DietNeRF を核に，不可視視点を MLLM によって推定する新たな枠組みを提案した．本手法により，MLLM の知識を誘導として活用することが，不可視視点のレンダリング表現の尤もらしさ向上に寄与することを確認した．

提案手法を用いることで，偏った視点からの 3 次元再構成のさらなる表現力向上が可能である．たとえば，PC 等の固定 Web カメラで正面のみ撮影された人物顔画像から，尤もらしい 3 次元再構成の生成への応用が期待される．

今後は，より尤もらしい不可視視点の補完を目指し，損失やその重みの改善を通して，前述の課題の解決を目指す．

## 参考文献


[1] Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S. et al.: BLIP3-o: A Family of Fully Open Unified Multimodal Models-Architecture, Training and Dataset, *arXiv preprint arXiv:2505.09568* (2025).

[2] Jain, A., Tancik, M. and Abbeel, P.: Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5885–5894 (2021).

[3] Kajiya, J. T. and Von Herzen, B. P.: Ray tracing volume densities, *ACM SIGGRAPH Computer Graphics*, Vol. 18, No. 3, pp. 165–174 (1984).

[4] Kerbl, B., Kopanas, G., Leimkühler, T. and Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering, *ACM Trans. Graph.*, Vol. 42, No. 4, pp. 139–1 (2023).

[5] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R. and Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, *Communications of the ACM*, Vol. 65, No. 1, pp. 99–106 (2021).

[6] Niemeyer, M., Barron, J. T., Mildenhall, B., Sajjadi, M. S., Geiger, A. and Radwan, N.: RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5480–5490 (2022).

[7] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al.: Learning Transferable Visual Models From Natural Language Supervision, *International Conference on Machine Learning*, PMLR, pp. 8748–8763 (2021).

[8] Schonberger, J. L. and Frahm, J.-M.: Structure-from-Motion Revisited, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4104–4113 (2016).

[9] Sun, Q., Fang, Y., Wu, L., Wang, X. and Cao, Y.: EVA-CLIP: Improved Training Techniques for CLIP at Scale, *arXiv preprint arXiv:2303.15389* (2023).

[10] Ultralytics: Explore Ultralytics YOLOv8, `https://docs.ultralytics.com/models/yolov8/` (2023). Accessed: 2025-06-15.

[11] Wang, Z., Li, J., Li, Y. and Liu, Y.: MutualNeRF: Improve the Performance of NeRF under Limited Samples with Mutual Information Theory, *arXiv preprint arXiv:2505.11386* (2025).

[12] Yang, J., Pavone, M. and Wang, Y.: FreeNeRF: Improving Few-shot Neural Rendering with Free Frequency Regularization, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8254–8263 (2023).