일반논문 (Regular Paper)
방송공학회논문지 제29권 제6호, 2024년 11월 (JBE Vol.29, No.6, November 2024)
https://doi.org/10.5909/JBE.2024.29.6.1056
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

단안 기하 단서를 이용한 가우시안-기반 사실적인 아바타 재구성

이 수 현[a], 김 서 연[a], 이 희 경[b], 정 원 식[b], 이 주 호[a)‡]

Gaussian-based Realistic Avatar Reconstruction With Monocular Geometric Priors

Soohyun Lee[a], Seoyeon Kim[a], Hee Kyung Lee[b], Won-Sik Jeong[b], and Joo Ho Lee[a)‡]

요 약

최근 Gaussian Splatting (GS)은 미분가능 레스터라이제이션을 통해 고충실도의 3차원 인간 아바타 재구성 및 실시간 렌더링을 가능케한다. 하지만 재구성된 Gaussian들이 실제 3차원 기하표면을 정확히 표현하지 못하고 다른 그래픽스 어플리케이션에 기하 오류를 야기한다. 우리는 GS 기반 아바타 재구성의 기하 정확성을 향상시키기 위해 단안 기하학적 단서를 최적화 과정에 활용한다. 구체적으로, 최신 단안 법선 및 깊이 추정 모델을 통해 이미지로부터 단안 기하 단서를 획득하고 이를 3차원 Gaussian들이 표현하는 3차원 표면에 강제하여 실제 표면과 근접하도록 한다. 본 연구에서는 단안 기하 단서의 유효성을 입증하고자 인간 단안시점 영상 데이터셋에서 재구성한 Gaussian 모델의 기하 정확성 및 렌더링 품질을 비교한다. 실험 결과, 단안 단서 조건항을 GS 아바타 재구성에 적용했을 때 렌더링 품질과 기하 정확성 모두에서 향상된 결과를 확인했다.

Abstract

Recently, Gaussian Splatting (GS) has been widely used in avatar reconstruction, achieving high-fidelity and real-time rendering by utilizing a differentiable rasterizer. Despite its remarkable performance, reconstructed Gaussians are often misaligned from the actual surface which leads to geometric errors. We propose to utilize monocular geometric cues in optimization in order to improve the geometric accuracy of GS-based human avatar reconstruction. We obtain monocular geometric cues from images using recent monocular depth and normal estimation models. The monocular geometric cues encourage 3D Gaussians to be aligned with the ground-truth surface. To prove the effectiveness of monocular geometric cues, we conduct the ablation study and measure geometric accuracy and rendering quality of 3D Gaussians reconstructed from monocular video human dataset. We demonstrate improvements in both rendering quality and geometric accuracy in GS avatar reconstruction with monocular geometric consistency term.

Keyword : Gaussian Splatting, Avatar Reconstruction, Monocular cues

Ⅰ. Introduction

Realistic avatar reconstruction has been intensively studied towards the seamless connection between the real and the virtual world. However, there is a trade-off between reconstruction quality and speed. Implicit function based approaches PIFu[11] achieve fast inference for 3D human avatars but require a huge training cost and large human scanning datasets, resulting in low geometric accuracy. In contrast, neural-based surface reconstruction NeRF[25] has been widely used in avatar reconstruction, especially for novel view synthesis, due to its outstanding rendering quality. Nevertheless, inverse skinning, fast reconstruction and real-time rendering are still challenging with neural representations.

Recently, Gaussian Splatting (GS)[2] has been proposed for addressing both reconstruction quality and speed in training and rendering. GS applied in avatar reconstruction is compatible with skinning methods , thanks to its explicit point-based representation. However, as noted in works like SuGaR[23] and 2DGS[24], GS suffer from misalignment between the reconstructed Gaussians and the ground-truth surface.

Inspired by MonoSDF[1], we propose a method that improves the geometric accuracy of Gaussians by utilizing monocular geometric cues in training. We align monocular depth maps to rendered depth maps by adjusting the scale and the depth offset using a least-square method. Then, we encourage the Gaussians to be aligned to the aligned monocular depth cue. We demonstrate that utilizing monocular cues on 3D human avatar reconstruction improves both the geometric and rendering quality.

a) 서강대학교 시각컴퓨팅연구실(Sogang University, Visual Computing Lab)
b) 한국전자통신연구원(Electronics and Telecommunications Research Institute)
‡ Corresponding Author : 이주호(Joo Ho Lee)
E-mail: jhleecs@sogang.ac.kr
Tel: +82-2-705-8489
ORCID: https://orcid.org/0000-0001-7307-7744
※ This work was partly supported by the immersive Media Research Laboratory (No. 2018-0-00207) and the Metaverse Convergence support program (IITP-2023-RS-2022-00156318), both supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP) funded by the Korea government (MSIT).
· Manuscript October 29, 2024; Revised November 11, 2024; Accepted November 11, 2024.

Ⅱ. Related Work

1. Monocular Geometric Cue Estimation

MonoSDF[1] demonstrates that the use of a general-purpose monocular estimator significantly improves both reconstruction quality and geometric accuracy for large scenes. Omnidata[3], used in [1], estimates monocular depth cues for a wide range of scenes. DN-Splatter[20] utilizes Omnidata and ZoeDepth[4] to enhance the geometry alignment in Gaussian Splatting using monocular depth and normal cues. ZoeDepth extends the relative depth prediction of MiDaS. Metric3Dv2[9] is a monocular estimation method based on Vision Transformer and ConvGRU. It addresses depth ambiguity by transforming images into the canonical camera space. To overcome the scarcity of surface normal datasets, Metric3Dv2 uses joint learning for depth and normal. DepthAnythingv2[8] employs a teacher-student framework, where the teacher is trained on synthetic datasets to circumvent the noise and incompleteness in real-world datasets. Recently, diffusion models like Marigold[6] and GeoWizard[7] have been proposed. Marigold, a depth estimation model, proposes training solely on synthetic datasets. GeoWizard introduces a cross-attention mechanism to ensure consistency between depth and normal predictions. However, diffusion-based models are known to be time-consuming.

On the other hand, ECON[13] introduces a normal integration method for reconstructing clothed humans from a single image. It estimates the front and back normal maps using a GAN-based model, as seen in [12], and this normal estimator can serve as a monocular cue. Additionally, the

recent model Sapiens[10], based on a Vision Transformer, offers broad applicability to various human-related tasks. To enhance generalizability, it uses the Human-300M dataset for pretraining, achieving state-of-the-art performance.

2. Gaussian splatting in Avatar Reconstruction

Gaussian splatting (GS) has been applied to various fields of 3D reconstruction due to its reconstruction quality along with fast training and rendering times. However, the reconstructed Gaussians are not aligned with the actual geometric surfaces. To address this, SuGaR[23] introduces regularization for Gaussian shapes and locations, while 2DGS[24] and GaussianSurfel[21] propose 2D Gaussians to resolve multi-view inconsistencies. Nevertheless, they do not address dynamic scenes or have not explored recent monocular estimators.

Some GS-based avatar reconstruction methods have been proposed. [15] uses pose features to decode Gaussian parameters for dynamic texture representation. [16] reconstruct a template by multi-view stereo then learns a network to estimate Gaussian maps from a position map at each time step. GART[14], on the other hand, employs a learnable skeleton to represent loose clothing. Though these models show high rendering quality, they still suffer from inaccurate 3D surfaces reconstructed by Gaussians. We improve 3D avatar reconstruction using monocular geometric cues. We compare different monocular geometric estimators in terms of depth and normal accuracy on human datasets.

III. Method

1. Preliminary: Gaussian Splatting

3D Gaussian Splatting (3DGS)[2] introduces volumetric 3D Gaussians to represent 3D scenes. Each 3D Gaussian represents a small cloud, having positional and radiometric properties such as mean position μ, rotation R, scale s, opacity o, and view dependent color c. Then, the Gaussian distribution at position x is parametrized by μ and 3D covariance matrix $\Sigma = RSS^TR^T$ as below:

$$G(x) = \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \quad (1)$$

The 3D Gaussian is splatted onto the image plane by the

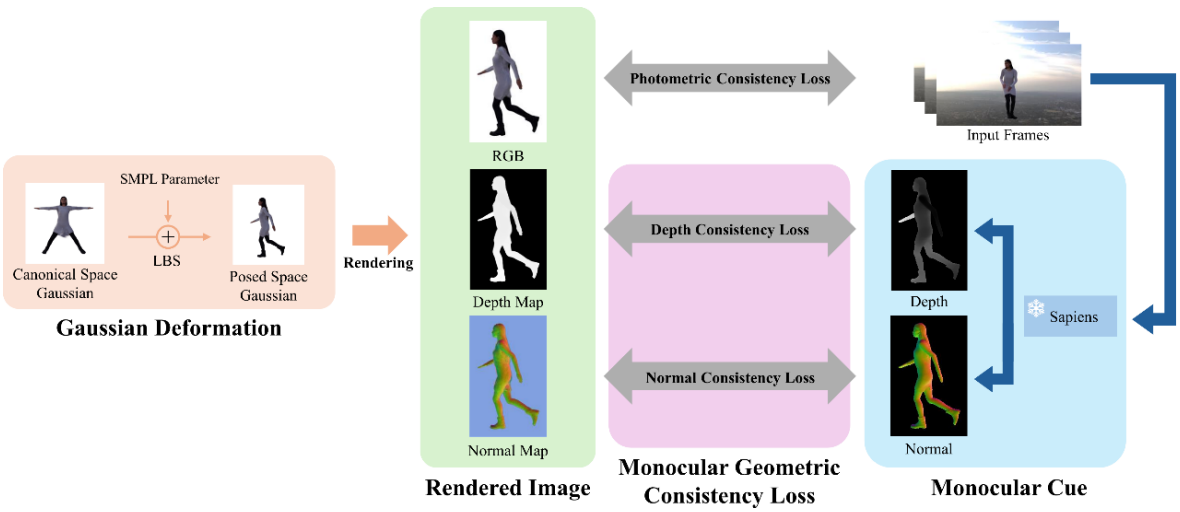

그림 1. 단안기하단서를 활용한 Gaussian Splatting 아바타 재구성 파이프라인 개요
Fig. 1. Overview of Gaussian Splatting avatar reconstruction using monocular geometric cues

elliptical weighted average (EWA) approximation since there is no analytic form to represent Gaussian projection. To accelerate the Gaussian rendering process and make it differentiable, the Gaussian renderer is implemented as the GPU-parallelized differentiable rasterizer. The rasterizer sorts Gaussians by depth from the view-point in fast tile-based approach. Then, Gaussians are blended with other Gaussians using the alpha-blending technique as shown in Equation 2:

$$c(x)=\sum_{i=1} c_i o_i G_i^{2D}(x)\prod_{j=1}^{i-1}\left(1-o_j G_j^{2D}(x)\right) \quad (2)$$

where $c(x)$ is color of a pixel $\mathbf{x}$, $G_i^{2D}(\mathbf{x})$ is a distribution of the ith Gaussian along ray $\mathbf{x}$, o_j and g_j are an opacity and a distribution of the jth intersected Gaussian before the ith Gaussian.

The goal of optimizing 3D Gaussians is to encourage rendered images I_{rend} aligned with ground-truth images I_{gt} so that Gaussians represent the appearance of the 3D scene. To this end, the difference between two images is computed as an objective function of the optimization process. This photometric loss L_{rgb} is computed as the combination of0 L_1 loss and $D-SSIM$ loss.

$$L_{rgb}=(1-\lambda_{SSIM})\left|I_{rend}-I_{gt}\right|+\lambda_{SSIM}SSIM(I_{rend},\ I_{gt}) \quad (3)$$

During optimization, overly-large Gaussians or sparsely-distributed Gaussians are struggling to reconstruct the surface and its texture. To alleviate this problem, 3DGS clone Gaussians in under-reconstructed regions and split large Gaussians in over-reconstructed regions.

2. Overview

We set GART[14], which is the state-of-the-art avatar reconstruction model by combining 3DGS with avatar reconstruction, as the baseline for our clothed human reconstruction. We first initialize Gaussians on the 3D body template mesh defined in the canonical space in Section 3.3. Then, Gaussians are deformed into the frame space by LBS in Section 3.4. and rendered into images. Our Gaussians learn the avatar's shape and appearance while minimizing the loss introduced in Section 3.5. The implementation details are described in Section 3.6.

3. Initialization

We define the canonical space of the avatar as the space with a "Da"-posed 3D body template as shown in Figure 1. "Da"-pose is a stance where the arms and legs are spread out wide to the both sides and it allows to reconstruct concave regions since a stretching-out stance reduces occlusions and overlaps between body parts. We locate Gaussians on the canonical mesh vertices. The orientation of each Gaussian aligns with normal of the mesh, scale is proportional to the face area, the opacity and the color are set to 0.9 and 0.5, respectively.

4. Deformation

Gaussians in the canonical space are animated according to the SMPL parameters θ_t at time t based on LBS. While the pose changes, the effects of key points on adjacent surfaces vary. We update the blending weight W_k of the kth key-point motion at each ith Gaussian by Δw:

$$\widehat{W}_k(\mu_i)=W_k(\mu_i)+\Delta w(\mu_i) \quad (4)$$

In clothed human reconstruction, it is challenging to describe cloth deformation by the traditional skinning method which represents deformation of a human skin. To account for dynamic cloth surfaces, we employ learnable latent bones $\tilde{B}(\theta_t)=\left[\tilde{B}_{t,1},...,\tilde{B}_{t,n_l}\right]$ to increase the expressivity of

the surface deformation model:

$$A_{t,i} = \sum_{k=1}^{n_b} \widehat{W}_k(\mu_i) B_{t,k} + \sum_{q=1}^{n_l} \widetilde{W}_q(\mu_i) \widetilde{B_{t,q}}, \quad (5)$$

$$\mu_{t,i}^{'} = A_{t,i} [\mu_i,\ 1]^T \quad (6)$$

where i indexes over Gaussians, $\mu_{t,i}'$ is a deformed Gaussian mean position to time t frame space, $\widetilde{W}$ is the blending weight of latent bones and $B(\theta_t) = [B_{t,1},...,B_{t,n_b}]$ are SMPL bones. n_b and n_l are number of SMPL bones and latent bones respectively. Then, Gaussians in the posed space are rendered into a RGB image, a normal map, and a depth map to compute loss and optimize Gaussians by following Section 3.5.

5. Optimization

We train Gaussians defined in the canonical space, their temporal deformation field for a given pose sequence during optimization. To make Gaussians of an avatar learn the appearance and 3D surface, we penalize the difference between rendered images and captured images and leverage the monocular geometric prior to enhance the surface details.

The photometric consistency loss L_{rgb} is the radiometric difference between the reconstructed video frames $I_{t,rend}$ $I_{t,rend}$ and observations $I_{t,rend}$:

$$L_{rgb} = \sum_t (1-\lambda_{SSIM}) |I_{t,\ rend} - I_{t,\ gt}| + \lambda_{SSIM} SSIM(I_{t,\ rend},\ I_{t,\ gt}), \quad (7)$$

where t is a time index. We penalize the photometric consistency loss to enhance the reconstruction fidelity.

The photometric consistency loss may not be sufficient to reconstruct smooth and natural surfaces. To resolve this issue, we leverage monocular geometric priors [1], which represent general surface shapes for given observation, in order to facilitate natural surface reconstruction. We encourage the reconstruction to have general surface orientations by fitting normal with monocular observation:

$$L_{norm} = \sum_t \| N_{t,rend} - N_{t,mono} \|_1 + \| 1 - N_{t,rend}^T N_{t,mono} \|. \quad (8)$$

where $N_{t,rend}$ is rendered normal and $N_{t,mono}$ is monocularly estimated normal.

We also utilize a monocular depth prior, representing the relative distance among pixels. We align the monocular depth map $D_{t,mono}$ to the reconstructed depth map $D_{t,rend}$ by scaling w and shifting q since there is the scale and shift ambiguity in monocular depth estimation:

$$L_{dep} = \sum_t \| (\omega D_{t,rend} + q) - D_{t,mono} \| \quad (9)$$

We encourage the smooth appearance by minimizing the standard deviation of Gaussian properties among K-nearest points:

$$L_{smooth} = \sum_{property \in \{R,s,o,\widehat{W},\widetilde{W}\}} \lambda_{property} STD_{i \in KNN(\mu_i)}(property_i) \quad (10)$$

where $property_i$ is a property of the ith Gaussian. In addition, We regularize non-rigid motion and size of Gaussians:

$$L_{scale,i} = \lambda_{\widehat{W}} \| \Delta w_i \| + \lambda_{\widehat{W}} \| \widetilde{W}(\mu_i) \| + \lambda_s \| s_i \|_\infty \quad (11)$$

where i means the ith Gaussian.

The total loss is sum of all loss terms for N Gaussians:

$$L_{total} = \lambda_{rgb} L_{rgb} + \lambda_{norm} L_{norm} + \lambda_{dep} L_{dep} + \frac{1}{N} \sum_{i=1}^{N} (L_{smooth,i} + L_{scale,i}) \quad (12)$$

6. Implementation Details

We jointly train Gaussian properties and motion parameters by applying the Adam optimizer with learning rates for each parameter the same as in [14], and other Adam hyperparameters are left at their default values of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We set λ for each loss term as $\lambda_{rgb} = 1.0$, $\lambda_{norm} = \lambda_{dep} = 0.05$ and $\lambda_{\tilde{W}}, \lambda_{\tilde{W}}, \lambda_s, \lambda_{property}$ the same as in [14]. The Gaussian parameters are optimized by stochastic gradient descent method for each time step t where t is randomly sampled from the training input video frame sequence. The entire framework is trained and tested on a single NVIDIA RTX A6000 GPU, with training times within 5 minutes and rendering times about 150 FPS.

Ⅳ. Experiments

1. Comparison between Monocular Estimators

Our Gaussian avatar learns surface of the object from monocular geometric cues. This makes the reconstruction quality of our method depend on monocular cue estimator performance. Thus, we analyze the state-of-the-art monocular normal and depth estimation models by qualitative and quantitative comparisons.

We evaluate the performance of normal and depth estimation of monocular estimators. We use the Thuman3.0[17] dataset, a real-world capture dataset of clothed humans which includes a variety of human and clothing scenes. It consists of 20 different clothed humans, each including between 15 to 35 pose sequences. For our evaluation, we select 5 poses from each clothed human. Since captured images are not provided, we synthesize multi-view images of scenes using the provided meshes and textures. For rendering, 8 point lights are placed at vertices of the bounding box surrounding the human mesh. We render four images from the front, back, and sides. The RGB images are used as input for the monocular estimation models, while rendered normal and depth maps are served as ground-truth data for evaluation.

1.1. Comparison on Surface Normal Estimation

We compare monocular estimators in terms of normal prediction. The evaluation metric “angular mean” represents the average angular distance, measured in degree, between rendered normals and the ground-truth. Lower values indicate better performance. The “ratio within x” metric represents the ratio of pixels whose angular distance with the ground-truth is less than x degrees. Higher ratio indicates higher performance. For Omnidata[3], we use two approaches proposed in DN-Splatter[20]: Omnidata-low and Omnidata-hd. The Omnidata-low approach resizes an input image, estimates normals, then resizes it back to the original resolution. The Omnidata-hd approach divides an input image into patches with overlaps and aggregates estimated normals of patches by aligning them with others.

In the case of Marigold[6], whose original version only estimate depth, we use the normal pretrained model available on Hugging Face. For Sapiens[10], we use the Sapiens-2B model trained with 2 billion parameters.

Table 1 shows that Omnidata-hd performs better than Omnidata-low. However, both Omnidata-based methods exhibit lower performance than other methods, indicating inadequacy of the model for human scenes. Vision Transformer-based models like Metric3Dv2[9] and Sapiens-2B outperform diffusion-based models like Marigold and GeoWizard[7], and ECON[13] demonstrates competitive results. GeoWizard, in particular, shows suboptimal performance in normal prediction. In contrast, Sapiens, which leverages pretraining on a large human-centric dataset, achieves the highest performance among the evaluated models.

Figure 2 illustrates the qualitative performance of mon-

표 1. 단안 표면 법선 예측 모델의 양적 비교 (빨간색: 1등, 파란색: 2등)
Table 1. Quantitative comparison of monocular normal estimation models (Red: 1st place, Blue: 2nd place)

	Omnidata-low	Omnidata-hd	ECON	Metric3Dv2	Marigold	Geowizard	Sapiens-2B
Angular Mean (degree °) ↓	31.3234	29.2753	22.4001	17.5526	20.0852	51.5194	13.2728
Ratio within 11.25° (%) ↑	11.1618	13.4923	26.8539	38.3021	31.1509	6.1990	57.0686
Ratio within 30° (%) ↑	54.2361	59.6652	76.7818	86.3450	80.9942	29.0239	92.4686

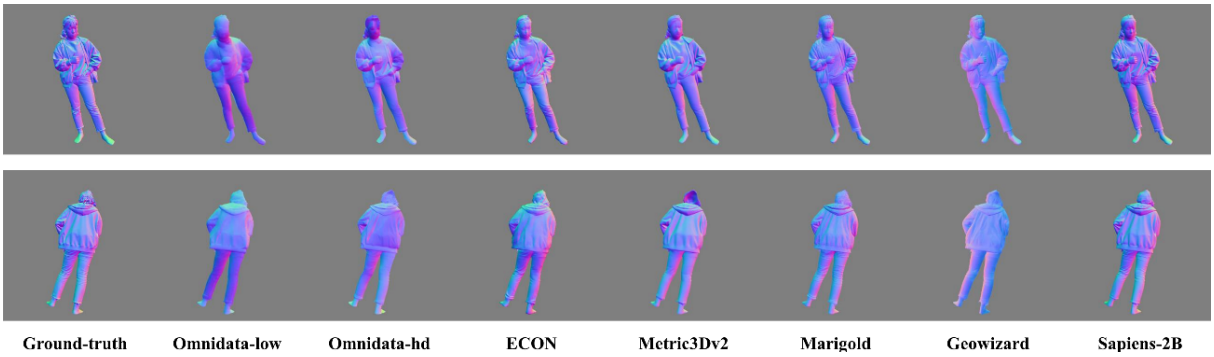

그림 2. 단안 표면 법선 예측 모델의 질적 비교
Fig 2. Qualitative comparison of monocular surface normal estimation models

ocular estimators. Omnidata-low lacks detail due to its resolution limitations, and while Omnidata-hd addresses alignment in the overlap regions, it still struggles with overall consistency. ECON captures large wrinkles well but fails to estimate precise normal direction. Marigold captures details effectively but lacks a sense of depth. Metric3Dv2 is highly accurate but suffers from inaccurate facial normals due to limited exposure to human scenes during training. Sapiens produces results close to the ground truth.

1.2 Comparison on Monocular Depth Estimation

We measure two depth metrics for the comparison of monocular estimators: AbsRel and δ^1. AbsRel represents the relative difference $|d-\hat{d}|/d$ between the ground-truth depth map d and the estimated depth map $\hat{d}$ with scale and shift adjustment. δ^1 represents the percentage of pixels whose ratio between the ground-truth and predicted values is below the threshold (1.25). We use the implementation of ZoeDepth[4] in DN-Splatter.

Table 2 demonstrates that Metric3Dv2 and Sapiens outperform other methods significantly. GeoWizard shows comparable performance to Marigold and ZoeDepth. Conversely, DepthAnythingv2[8] produces results that deviates significantly from the ground truth. Overall, Sapiens demonstrates the most realistic results.

표 2. 단안 깊이 예측 모델의 양적 비교 (빨간색: 1등, 파란색: 2등)
Table 2. Quantitative comparison of monocular depth estimation models (Red: 1st place, Blue: 2nd place)

	ZoeDepth	DepthAnythingv2	Metric3Dv2	Marigold	Geowizard	Sapiens-2B
AbsRel ↓	0.0321	0.1745	0.0197	0.0357	0.0327	0.0129
δ^1 ↑	0.9989	0.7008	0.9999	0.9958	0.9937	1.0000

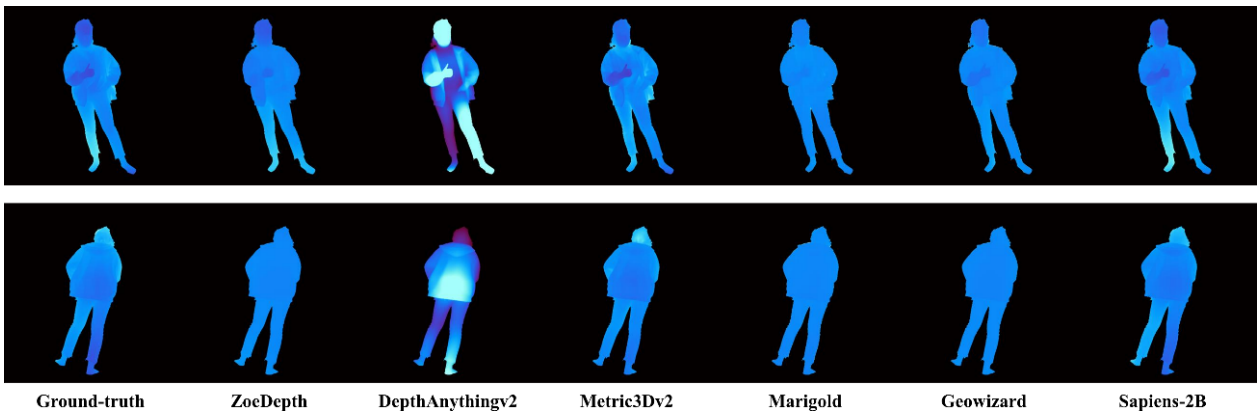

그림 3. 단안 깊이 예측 모델의 질적 비교
Fig. 3. Qualitative comparison of monocular depth estimation models

Figure 3 illustrates qualitative performance of monocular estimators. DepthAnythingv2 fails to align the scale and shift correctly due to the large difference between the minimum and maximum depth scales. ZoeDepth, Marigold, and GeoWizard appear to lack high-frequency details. In contrast, Metric3Dv2 and Sapiens produce results that closely approximates the ground truth.

2. Avatar Reconstruction

We evaluate the geometric quality of avatar reconstruction. We use the RANA[18] dataset, a photorealistic synthetic dataset of clothed humans that provides normal maps, camera parameters, and SMPL[19] parameters. We select 5 subjects for evaluation. Each subject contains 150 frames, with the first 100 frames used for Gaussian optimization and the remaining 50 frames reserved for testing.

We conduct qualitative and quantitative comparison. We use Sapiens[10] as a model for monocular geometric consistency due to its superior performance demonstrated in Sections 4.1.2 and 4.1.3. Table 3 quantitatively verifies that the monocular geometric consistency constraint improves both rendering quality and geometric quality. Specifically, after the GS process, the accuracy of the normal map is significantly improved.

Figure 4 shows the qualitative comparison between the baseline and reconstruction with monocular consistency constraint. The normal map reconstructed with the monocular consistency prior is smoother, aligned to the ground-truth, and reducing noise in the texture of the rendered RGB images. Notably, it is challenging for the Gaussian to reconstruct wrinkles in clothing, but Sapiens' precise normal and depth estimation plays a key role in improving the geometric quality of the clothing, especially in capturing the folds.

표 3. 단안 단서를 활용한 Gaussian Splatting 아바타 재구성 방식의 양적 비교
Table 3. Quantitative comparison of Gaussian Splatting avatar reconstruction using monocular cues

Sequence	Method	PSNR ↑	SSIM ↑	LPIPS ↓	Angular Mean (°) ↓	% within 11.25° ↑	% within 30° ↑
Subject_31	GART	30.8105	0.9847	0.0117	45.6566	2.0187	25.1155
	Ours	31.7329	0.9864	0.0137	42.9615	1.9636	27.2319
Subject_41	GART	29.3117	0.9832	0.0188	46.0633	2.2215	25.2712
	Ours	30.3213	0.9848	0.0206	41.7219	2.2367	33.2051

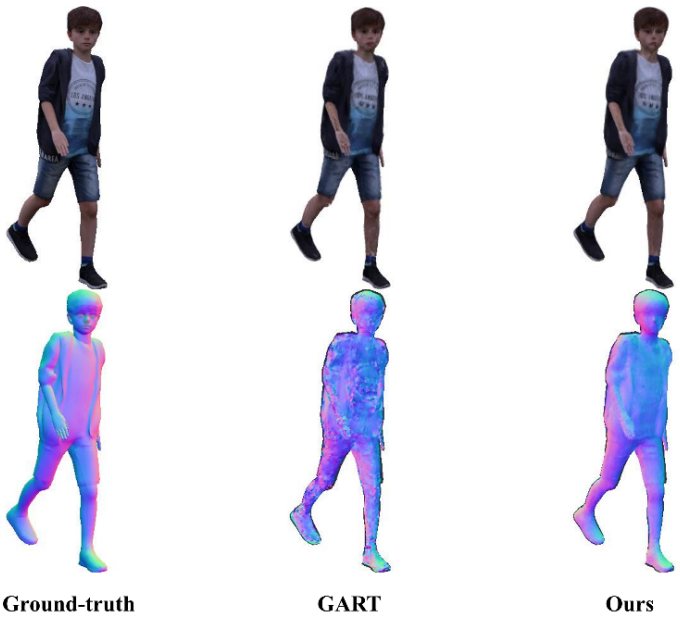

그림 4. 단안 단서를 활용한 Gaussian Splatting 아바타 재구성 방식의 질적 비교
Fig. 4. Qualitative comparison of Gaussian Splatting avatar reconstruction using monocular cues

Ⅴ. Conclusion

In this work, we propose to utilize monocular geometric cues to enhance surface alignment and geometric accuracy of reconstructed Gaussians in GS avatar reconstruction. To this end, we compare the performance of various monocular estimation models in human scanning dataset. We conclude that Sapiens, pretrained on enormous collection of human datasets, is the most powerful tool for human-centric tasks. Thus, we apply this model to GS avatar reconstruction and we demonstrate that monocular geometry cue estimated by Sapiens leads to smoother and aligned texture and normal of reconstructed Gaussians.

참 고 문 헌 (References)

[1] Y Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, “Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction”, Advances in neural information processing systems 35, pp. 25018-25032, 2022.
doi: https://dl.acm.org/doi/10.5555/3600270.3602084

[2] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, “3D Gaussian Splatting for Real-Time Radiance Field Rendering”, ACM Transactions on Graphics (TOG), Vol.42, No.4, pp. 1-14, July 2023.
doi: https://doi.org/10.1145/3592433

[3] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, “Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans”, In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10786-10796, 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.01061

[4] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, “Zoedepth: Zero-shot transfer by combining relative and metric depth”, ArXiv preprint ArXiv:2302.12288, 2023.
doi: https://doi.org/10.48550/arXiv.2302.12288

[5] W. Chen, S. Qian, D. Fan, N. Kojima, M. Hamilton, and J. Deng, “Oasis: A large-scale dataset for single image 3d in the wild”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 679-688, 2020.
doi: https://doi.org/10.1109/CVPR42600.2020.00076

[6] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, “Repurposing diffusion-based image generators for monocular depth estimation”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9492-9502, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.00907

[7] X. Fu, W. Yin, M. Hu, K. Wang, Y. Ma, P. Tan, S. Shen, D. Lin, and X. Long, “Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image”, In European Conference on Computer Vision, pp. 241-258, September 2024.

doi: https://doi.org/10.1007/978-3-031-72670-5_14

[8] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, “Depth Anything V2”, ArXiv Preprint ArXiv: 2406.09414, 2024.
doi: https://doi.org/10.48550/arXiv.2406.09414

[9] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, and S. Shen, “Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-shot Metric Depth and Surface Normal Estimation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 46, No.12, pp. 10579-10596, December 2024.
doi: https://doi.org/10.1109/TPAMI.2024.3444912

[10] R. Khirodkar, T. Bagautdinov, J. Martinez, S. Zhaoen, A. James, P. Selednik, S. Anderson, and S. Saito, “Sapiens: Foundation for Human Vision Models”, In European Conference on Computer Vision, pp. 206-228, September 2024.
doi: https://doi.org/10.1007/978-3-031-73235-5_12

[11] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, “Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization”, In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2304-2314, 2019.
doi: https://doi.org/10.1109/ICCV.2019.00239

[12] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black, “Icon: Implicit clothed humans obtained from normals”, In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13286-13296, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.01294

[13] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black, “Econ: Explicit clothed humans optimized via normal integration”, In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 512-523, 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.00057

[14] J. Lei, Y. Wang, G. Pavlakos, L. Liu, and K. Daniilidis, “Gart: Gaussian articulated template models”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19876-19887, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.01879

[15] L. Hu, H. Zhang, Y. Zhang, B. Zhou, B. Liu, S. Zhang, and L. Nie, “Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 634-644, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.00067

[16] Z. Li, Z. Zheng, L. Wang, and Y. Liu, “Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19711-19722, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.01864

[17] Z. Su, T. Yu, Y. Wang, and Y. Liu, “Deepcloth: Neural garment representation for shape and style editing”, IEEE Transactions on Pattern Analysis and Machine Intelligence 45.2, pp. 1581-1593, 2022.
doi: https://doi.org/10.1109/TPAMI.2022.3168569

[18] U. Iqbal, A. Caliskan, K. Nagano, S. Khamis, P. Molchanov, and J. Kautz, “Rana: Relightable articulated neural avatars”, In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 23142-23153, 2023.
doi: https://doi.org/10.1109/ICCV51070.2023.02115

[19] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M. J. Black, “SMPL: A Skinned Multi-Person Linear Model”, In ACM Transactions on Graphics, pp. 1-16, 2015.
doi: https://doi.org/10.1145/2816795.2818013

[20] M. Turkulainen, X. Ren, I. Melekhov, O. Seiskari, E. Rahtu, and J. Kannala, “DN-Splatter: Depth and Normal Priors for Gaussian Splatting and Meshing.”, ArXiv Preprint ArXiv: 2403.17822, 2024.
doi: https://doi.org/10.48550/arXiv.2403.17822

[21] P. Dai, J. Xu, W. Xie, X. Liu, H. Wang, W. Xu, “High-quality surface reconstruction using gaussian surfels”, In ACM SIGGRAPH 2024 Conference Papers, pp. 1-11, 2024.
doi: https://doi.org/10.1145/3641519.3657441

[22] M. Zwicker, H. Pfister, J. V. Baar, and M. Gross, “EWA volume splatting”, In Proceedings Visualization, VIS’01, IEEE, pp. 29-538, 2001.
doi: https://doi.org/10.1109/VISUAL.2001.964490

[23] A. Guédon, and V. Lepetit, “Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5354-5363, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.00512

[24] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao, “2d gaussian splatting for geometrically accurate radiance fields”, In ACM SIGGRAPH 2024 Conference Papers, pp. 1-11, 2024.
doi: https://doi.org/10.1145/3641519.3657428

[25] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis”, Communications of the ACM, 65(1), pp. 99-106, 2021.
doi: https://doi.org/10.1145/3503250

[26] S. Wang, B. Antic, A. Geiger, and S. Tang, “IntrinsicAvatar: Physically Based Inverse Rendering of Dynamic Humans from Monocular Videos via Explicit Ray Tracing”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1877-1888, 2024.
doi: http://dx.doi.org/10.1109/CVPR52733.2024.00184

저 자 소 개

이 수 현

- 2023년 : 서강대학교 수학/컴퓨터공학(학사)
- 현재 : 서강대학교 인공지능학 석사과정
- ORCID : https://orcid.org/0009-0000-8158-0622
- 주관심분야 : 컴퓨터비전, 컴퓨터 그래픽스

김 서 연

- 2024년 : 서강대학교 컴퓨터공학(학사)
- 현재 : 서강대학교 인공지능학 석사과정
- ORCID : https://orcid.org/0009-0001-4977-0810
- 주관심분야 : 컴퓨터비전, 컴퓨터 그래픽스

이 희 경

- 1999년 : 영남대학교 컴퓨터공학과(공학사)
- 2022년 : 한국정보통신대학교(ICU) 공학부(공학석사)
- 2002년 ~ 현재 : 한국전자통신연구원 책임연구원
- ORCID : https://orcid.org/0000-0002-1502-561X
- 주관심분야 : 컴퓨터 비전, 기계학습, VCM, 메타버스, 360VR, 시선추적, 메타데이터

정 원 식

- 1992년 : 경북대학교 전자공학과(공학사)
- 1994년 : 경북대학교 대학원 전자공학과(공학석사)
- 2000년 : 경북대학교 대학원 전자공학과(공학박사)
- 2000년 ~ 현재 : 한국전자통신연구원 책임연구원
- ORCID : https://orcid.org/0000-0001-5430-2969
- 주관심분야 : 이머시브 미디어 기술, 기계를 위한 영상 부호화, 딥러닝기반 신호처리, 멀티미디어 표준화

이 주 호

- 2012년 : 한국과학기술원 전산학과(학사)
- 2014년 : 한국과학기술원 전산학부(석사)
- 2020년 : 한국과학기술원 전산학부(박사)
- 현재 : 서강대학교 컴퓨터공학과 교수
- ORCID : https://orcid.org/0000-0001-7307-7744
- 주관심분야 : 컴퓨터 그래픽스, 컴퓨터 비전, 시각 컴퓨팅, 3차원 재구성