TimeWalker: Personalized Neural Space for Lifelong Head Avatars

Dongwei Pan[1] , Yang Li[1] , Hongsheng Li[2] , Kwan-Yee Lin[1]

[1] Shanghai AI Laboratory [2] CUHK

linjunyi9335@gmail.com

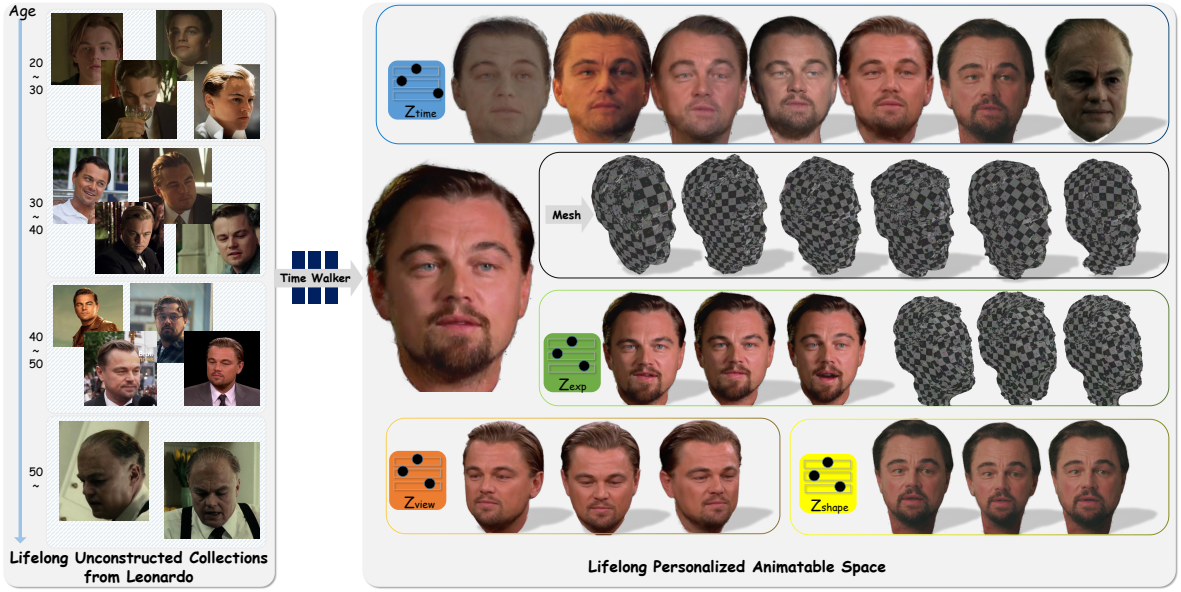

Figure 1. **TimeWalker.** Given a set of unstructured data from the Internet or photo collection across years, we build a personalized neural parametric morphable model, *TimeWalker*, towards replicating a life-long 3D head avatar of a person. With the TimeWalker, we can control and animate one's avatar in terms of shape, expression, viewpoint, and appearance across his/her different age periods. In this Figure, We show Leonardo Dicaprio's life-long avatar reconstructed and animated by our proposed model.

Abstract

*We present **TimeWalker**, a new framework that models realistic, full-scale 3D head avatars of a person on life-long scale. Unlike current human head avatar pipelines that capture a person's identity only at the momentary level (i.e., instant photography, or short videos), TimeWalker constructs a person's comprehensive identity from unstructured data collection over his/her various life stages, offering a paradigm to achieve full reconstruction and animation of that person at different moments of life. At the heart of TimeWalker's success is a novel neural parametric model that learns personalized representation with the disentanglement of shape, expression, and appearance across ages. Central to our methodology are the concepts of two aspects:* (1) *We track back to the principle of modeling a person's identity in an additive combination of his/her average head representation in the canonical space, and moment-specific head attribute representations driven from a set of neural head basis. To learn the set of head basis that could represent the comprehensive head variations of the target person in a compact manner, we propose a **Dy**namic **N**eural **Ba**sis-Blending **Mo**dule (**Dynamo**). It dynamically adjusts the number and blend weights of neural head bases, according to both shared and specific traits of the target person over ages.* (2) *We introduce **Dyna**mic **2**D Gaussian **S**platting (**DNA-2DGS**), an extension of Gaussian splatting representation, to model head motion deformations like facial expressions without losing the realism of rendering and reconstruction of full head. DNA-2DGS includes a set of controllable 2D oriented planar Gaussian disks that utilize the priors from a parametric morphable face model, and move/rotate with the change of expression. Through extensive experimental evaluations, we show TimeWalker's abil-*

ity to reconstruct and animate avatars across decoupled dimensions with realistic rendering effects, demonstrating a way to achieve personalized "time traveling" in a breeze.

1. Method

Our pipeline aims to construct a comprehensive head avatar of a person's identity across their different life stages, as opposed to current head avatar methods that reconstruct and animate a person at the momentary level. The main challenge of constructing the head avatar on a lifelong scale is the additional embedding of the lifestage dimension during the modelling process. The changes to the head brought about by the different lifestage of a person cannot be explicitly defined, as it involves differences in appearance, facial shape or even accessories, while at the same time it has to be disentangled from the other dimension to enable the decoupled animation. To address the challenges, we introduce a novel neural parametric model that models the average representation of a person's identity in the canonical space with the form of set of 2D Gaussian Surfels [2] and spans to moment-specific head attribute representations by driving a set of Neural Head Basis. Further, we extend the 2DGS [2] to DNA-2DGS module, a dynamic version to reconstruct and drive the dense mesh with different motion signals.

Neural Head Basis. Building on the concept of traditional 3D Morphable Head Models [1, 5], we introduce the Neural Head Basis, which efficiently captures moment-specific features of an individual's head. To store these features compactly, we utilize a Multi-resolution Hashgrid [6], a hashmap-based cubic structure, which enables the storage of learnable features in a condensed form. When given a canonical point location $\mathbf{x}_c$ from Gaussian kernels, the hashgrid lookup nearby features at various scales, and cubic linear interpolation is applied to obtain the final feature corresponding to the location. By employing multiple hashgrids $\{\mathcal{H}_i\}_{i=1}^{N}$, we encompass a comprehensive range of head variations in our model, including both common features and those specific to individual's different life stages. Our goal is to ensure that our neural basis learns these deeper characteristics rather than solely memorizing superficial appearances. To this end, we introduce the Dynamic Neural Basis-Blending module (**Dynamo**), which dynamically adjusts the number of basis during during the learning process of blending weight and hashgrid. Specifically, we initialize a set of learnable blending weights $\{\beta\}_{i=1}^{N}$ and perform a weighted sum of the features extracted from the neural basis: $\mathbf{f}(\mathbf{x}_c) = \sum_{i=1}^{N} \beta_i \mathcal{H}_i(\mathbf{x}_c)$. Throughout the learning process, we continuously monitor the blending weights of the N hashgrids. If a hashgrid's weight is consistently low across data of multiple lifestages, it indicates that the grid is not effectively learning the character's features. In response, we deactivate that particular hashgrid. By the end of the training phase, we can guarantee that all Neural Head Basis are actively learning valid appearance features, including deep invariant characteristics of the character. The features from Dynamo are then feed into compact MLPs to derive the attributes of the Gaussian kernel. The network learns the deformations of the Gaussian attributes, which are then additively combined with the Gaussian average in canonical space.

Gaussian Surfels Representation. To characterise the average head representation of individuals, we define a set of Gaussian surfels in canonical space, initially positioned on the face of the FLAME template [5]. With addtion of the deformation value produced from Neural Head Basis, the canonical Gaussian Surfels are deformed from the average representation to a specific lifestage. Gaussian Surfels after deformation perform appearance and underlining characteristic in a static way. Afterwards, to add motion and realize dynamic avatar animation, we utilize the motion warping fields rooted from FLAME [5] expression & shape parameters. This two deformation guidance, the deformation fields that drives the mean representation to moment-specific static avatar, and motion warping fields that empowers the head model with dynamic motion, allows us to create multi-dimensional realistic head avatar.

DNA-2DGS. The existing Gaussian Surfels [2] technique allows for the reconstruction of high-quality surfaces after training through Gaussian point cutting and Poisson meshing [3] with extracted data from rendering results. However, this approach is primarily suitable for static scenes and cannot be directly applied to dynamic head avatars. Furthermore, the Poisson reconstruction process employed in this method is time-consuming, making it impractical to reconstruct each frame individually. To address these limitations, we propose an adaptation and extension of the Gaussian surface reconstruction method specifically tailored for dynamic head reconstruction. Specifically, we do not perform data extraction on the rendering results in deformed space, but rather we take a step back and render the results under moment-specific conditions, tailored to different appearance data. This process effectively removes any interference caused by motion, enabling us to obtain appearance-specific static mesh. Unlike warping the Gaussian surfels during the rendering process, here we warp the vertices of the reconstructed mesh. By doing so, we can generate motion-driven mesh results that accurately capture the dynamics of the subject. Notably, the above scheme, we call Defer-Warping, allows us to obtain dynamic mesh sequences in a much shorter time comsume.

Training. We apply the end-to-end training manner that enables the simultaneous optimization of the explicit Gaussian surfels, multiple hashgrid, and implicit MLPs. For the Gaussian Splatting, we follow the densify and pruning strategies of 3DGS to adaptively adjust the number of

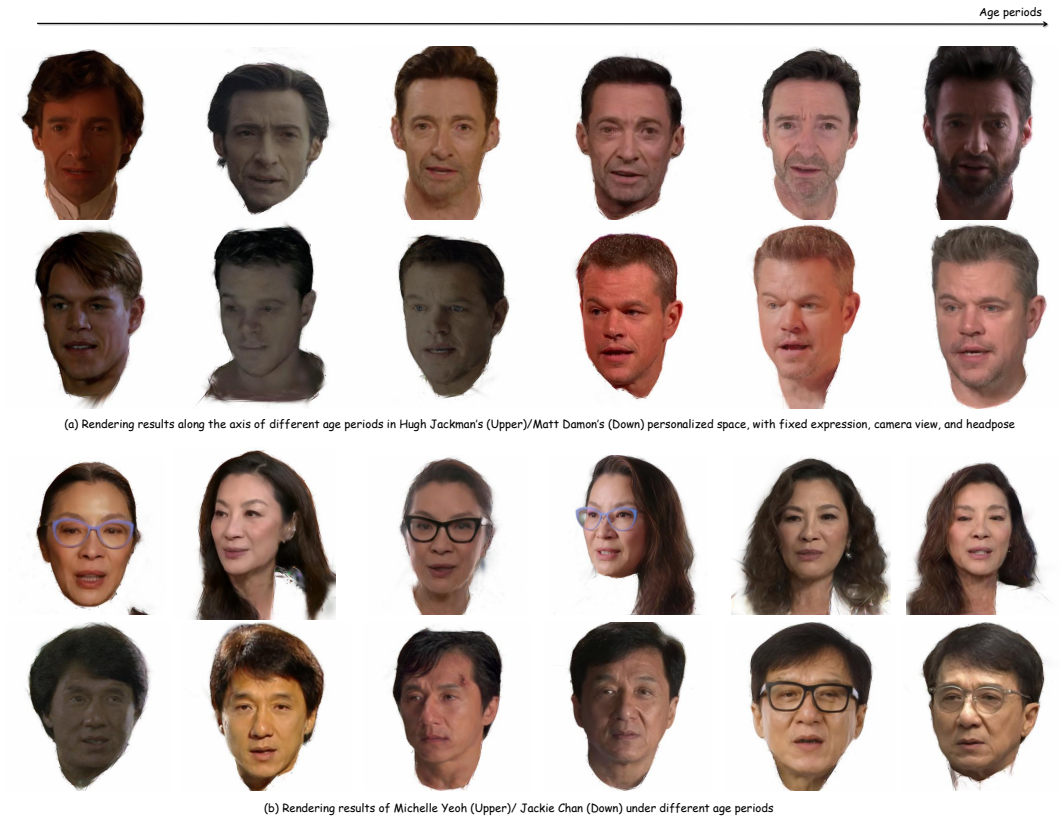

Figure 2. **Personalized Space: Lifestage.** We demonstrate multiple individuals and their replicas in different lifestages. (a) We adjust the value in Dynamo to animate the lifestages of individuals, but keep other animation values unchanged. (b) We show the lifestages of more individuals with another ethnicity.

Gaussian. To guide the optimization of the whole system, our total loss $\mathcal{L}_{\text{total}}$ consists of three parts: (1) Image Level Supervision. Similar to 3DGS [4], This term includes photometric $L1$ loss $\mathcal{L}_{\text{rgb}}$ and ssim loss $\mathcal{L}_{\text{ssim}}$. (2) Geometry Level Supervision. We introduce $\mathcal{L}_{\text{depth}}$ from INSTA [8] to enforce a better Gaussian geometry, and $\mathcal{L}_{\text{normal}}$ from Guassian Surfels [2]. (3) Regulation. To avoid the Gaussian attribute not walking far from its average representation, we employ a $L1$ regulation to the deformation of the Gaussian attributes and penalize large deformation.

Building a Life Long Personalized Space. We decouple the driving of the head in multiple dimensions - lifestage, expression, shape and novel view, - as described in the following – (1) Lifestage: During training, our pipeline learns different blending weights for data in different life stages. After training, we can adjust these weights to drive the lifestage in a disentangled manner. Fig. 2 illustrates the appearance diversity of individuals as they progress through different life stages. This demonstrates the effectiveness of our pipeline in capturing a person's identity across different moments in their life. (2) Expression: To achieve expression and shape changes of the character while maintaining a consistent appearance, we use a motion warping field inspired from INSTA [8]. By manipulating expression parameters, we can update both the tracked mesh and the transformation matrix that maps from canonical space to deformation space. This enables us to achieve the desired expression-based warping. (3) Shape: As the FLAME mesh can be driven by expression and shape parameters in a disentangle manner, our head avatar can also be animated by shape with the same approach as expression. (4) Novel view: The Gaussian Splatting, as type of 3D representation, can be rendered with arbitrary camera pose.

2. Experiments

	PSNR↑	SSIM↑	LPIPS↓
w/o $\mathcal{L}_{\text{lpips}}$	26.56	0.916	0.18
w/o $\mathcal{L}_{\text{geometry}}$	27.25	0.943	0.080
w/o $\mathcal{L}_{\text{deform}}$	24.79	0.886	0.165
w/o Dynamo	21.69	0.767	0.197
w/ 1 hashgrid	24.84	0.890	0.119
w/ all hashgrid	26.86	0.938	0.078
Ours	27.20	0.941	0.077

Table 1. Ablation study. Pink indicates the best and orange indicates the second.

Dataset. To fully validate the effectiveness of our pipeline, we construct a large-scale head dataset, named

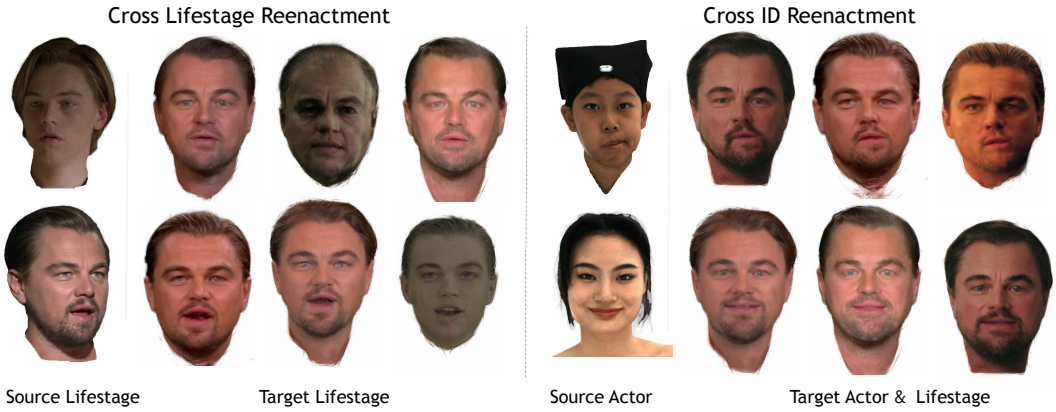

Figure 3. **Self Reenactment.** We demonstrate the cross-lifestage reenactment with TimeWalker-1.0 and cross-identity reenactment with RenderMe-360 [7], in Leonardo personalized space.

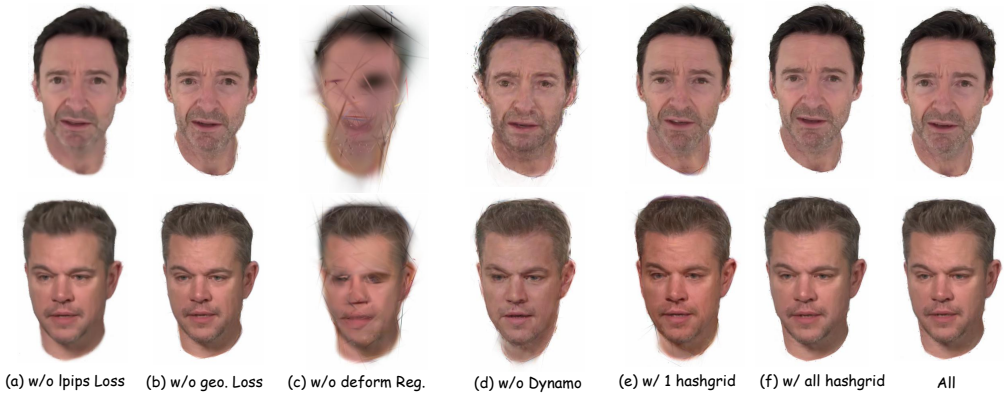

Figure 4. **Ablation Study.** Experiments with different loss setting are showed in $(a-c)$, while ablation with Dynamo and hashgrid are visualized in $(d-f)$.

TimeWalker-1.0., which includes 20 celebrities' lifelong photo collections. The data volume ranges from 10K to 100K for each celebrity, with diverse variations over different lifestages (e.g., shape, headpose, expression, and appearance).

Reenactment. Fig. 3 performs expression animation through two types of reenactments. The cross-lifestage reenactment on the left showcases how head avatars from different life stages can consistently perform the same expression, animated from a source avatar belonging to a different time period. In the right part Leonardos in different lifestages are driven by unseen novel expression from RenderMe-360 [7], and the rendering result shows that multiple head avatars generated by the same personal space from TimeWalker are able to extrapolate novel expression.

Ablation Studies. To validate the effectiveness of our method components, we conduct several ablation experiments in terms of our Dynamo design and loss terms. All the ablation experiments are conducted on 3 individuals with each at least 9 lifestages. We keep other settings unchanged except the ablation term, whose results are demonstrated in Tab.1 and Fig. 4.

3. Conclusion

Limitations. Firstly, our personalized avatar fails to capture exaggerated expressions during the animation, resulting in noticeable artifacts around the mouth area. Secondly, when rendering avatars with thin structures from a large novel view, significant blurring occurs. We owe these two limitations to the limited diversity of the expression and head pose within the data from one life stage. Regardless of the data source, a promising approach is to harness the prior knowledge of head structure from pre-trained generative models, which we leave as a direction for future exploration.

Conclusion. In this work, we present the TimeWalker, a baseline solution to construct a personalized space with long-horizon identity consistency preserving and explicit-controlled animation in full scales. The key design lies in the two components: a Dynamic Neural Basis-Blending model to represent the head variations in a compact manner and a Dynamic 2D Gaussian Splatting module to construct dynamic dense head mesh. Our methods are capable of building personalized space on a life-long scale.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *CGIT*, 1999. 2

[2] Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using gaussian surfels. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. 2, 3

[3] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 2

[4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 3

[5] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *TOG*, 2017. 2

[6] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *TOG*, 2022. 2

[7] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, et al. Renderme-360: A large digital asset library and benchmarks towards high-fidelity head avatars. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[8] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4574–4584, 2023. 3