From Gallery to Wrist: Realistic 3D Bracelet Insertion in Videos

Supplementary Material

We highly recommend that readers view the **supplementary videos** provided alongside this document to explore additional results and visualizations. Below, we include technical details omitted from the main text.

1. Bracelet 3D Motion Calculation

Given a 3D Gaussian Splatting (3DGS) model $\mathcal{G}$ of the bracelet and its initial pose $\mathbf{P}_1 = (\mathbf{R}_1, \mathbf{T}_1)$ in the first frame, we compute the bracelet's pose in subsequent frames to align it with the wrist's motion. We rely on 2D keypoint tracking result of skin near the bracelet, and use them to calculate 3D bracelet motion.

Specifically, this is modeled as a *Perspective-n-Point (PnP)* [1] problem. The input is a set of 2D points $\{\mathbf{x}_t^i\}_{i=1}^N$ on the skin, initialized in the first frame as $\{\mathbf{x}_1^i\}_{i=1}^N$, and tracked across frames using *CoTracker* [2]. Then, we lift these 2D points $\{\mathbf{x}_1^i\}$ to 3D using the depth map $\mathbf{D}_1$, camera intrinsics $\mathbf{K}$, and the initial pose $\mathbf{P}_1$:

$$\mathbf{X}_1^i = \mathbf{P}_1^{-1} \cdot \left(\mathbf{K}^{-1} \begin{bmatrix} \mathbf{x}_1^i \\ \mathbf{D}_1(\mathbf{x}_1^i) \end{bmatrix} \right),$$

where $\mathbf{D}_1$ and $\mathbf{K}$ are estimated by *UniDepth* [5]. For each frame t, the corresponding pose is computed by solving PnP equation:

$$\arg\min_{\mathbf{P}_t} \sum_{i=1}^{N} \|\mathbf{K}\mathbf{P}_t\mathbf{X}_1^i - \mathbf{x}_t^i\|^2.$$

Finally, we smoothed the poses by applying a bilateral filter to the translation vector $\mathbf{T}$ and quaternion representations of the rotation $\mathbf{R}$ to get the final $\mathbf{P}_t$.

2. Occlusion Handling

To ensure a correct depth ordering between human, background, and bracelet, we use monocular depth maps $\mathbf{D}_t$ from *UniDepth* as the 3D context to handle the occlusion. To align depth across frames, we compute a scale factor s_t for each frame t using the calculated pose $\mathbf{P}_t$:

$$\arg\min_{s_t} \sum_{i=1}^{N} \|s_t \cdot \mathbf{D}_t(\mathbf{x}_t^i) - [\mathbf{P}_t\mathbf{X}_1^i]_z\|^2.$$

The solution to this optimization problem is:

$$s_t = \frac{\sum_{i=1}^{N} \mathbf{D}_t(\mathbf{x}_t^i) \cdot [\mathbf{P}_t\mathbf{X}_1^i]_z}{\sum_{i=1}^{N} \mathbf{D}_t(\mathbf{x}_t^i)^2}.$$

Once the depth maps are aligned, we compute the occlusion mask $\mathbf{M}_t$ for each frame by comparing the scene depth $\mathbf{D}_t$ with the bracelet's depth rendering $\mathbf{D}_t^{\text{bracelet}}$.

$$\mathbf{O}_t(\mathbf{x}) = \begin{cases} 1 & \text{if } s_t \cdot \mathbf{D}_t(\mathbf{x}) < \mathbf{D}_t^{\text{bracelet}}(\mathbf{x}), \\ 0 & \text{otherwise.} \end{cases}$$

To avoid hard edges and aliasing artifacts, we further apply a Gaussian blur to the initial occlusion map $\mathbf{O}_t$, yielding the soft mask $\mathbf{M}_t = G_\sigma * \mathbf{O}_t$, where G_σ is a Gaussian kernel with standard deviation σ.

Finally we compute a preview of the bracelet in the current frame:

$$\mathbf{I}_t^{\text{preview}}(\mathbf{x}) = \mathbf{M}_t(\mathbf{x}) \cdot \mathbf{I}_t^{\text{bracelet}}(\mathbf{x}) + (1 - \mathbf{M}_t(\mathbf{x})) \cdot \mathbf{I}_t^{\text{scene}}(\mathbf{x}),$$

where $\mathbf{I}_t^{\text{bracelet}}$ is the rendered image of the bracelet at pose $\mathbf{P}_t$, and $\mathbf{I}_t^{\text{scene}}$ is the original scene image. This preview is an intermediate output, with further refinements applied later to enhance realism.

3. Optimizing 3D Gaussian for Smoothing

To achieve smooth temporal transitions while preserving the geometric structure of the bracelet, we optimize only the spherical harmonics (SH) coefficients associated with each splat in the 3D Gaussian Splatting (3DGS) model [3]. The SH coefficients encode view-dependent color information for each Gaussian splat, allowing us to adjust the appearance of the bracelet without modifying its geometric attributes (position, scale, rotation, and opacity). For each splat i, let $\mathbf{c}_i(\mathbf{d})$ represent its view-dependent color, which is a function of the viewing direction $\mathbf{d}$. This color is computed using the spherical harmonics basis functions $Y_l^m(\mathbf{d})$ and the corresponding SH coefficients $\mathbf{s}_i$, where $s_{i,l}^m$ are the SH coefficients for splat i, with l and m representing the degree and order of the spherical harmonics, respectively, $Y_l^m(\mathbf{d})$ are the spherical harmonics basis functions evaluated at the viewing direction $\mathbf{d}$.

During optimization, we update the SH coefficients $\mathbf{s}_i$ for each splat i to minimize the photometric error between the rendered image and the refined reference image $\mathbf{I}_t^{\text{refined}}$. This ensures that the color and shading of the bracelet are smoothly adjusted while keeping the geometric structure fixed. The optimization can be expressed as:

$$\mathbf{s}_i^* = \underset{\mathbf{s}_i}{\text{argmin}} \sum_{k=t-W/2}^{t+W/2} w(k-t) \cdot \|\mathcal{R}(\mathbf{K}, \mathbf{P}_k, \mathcal{G}(\mathbf{s}_i)) - \mathbf{I}_t^{\text{refined}}\|^2,$$

where $\mathcal{G}(\mathbf{s}_i)$ denotes the 3DGS model with updated SH coefficients $\mathbf{s}_i$ for each splat. We employ the Adam [4] optimizer to solve this optimization problem, with a learning rate of 10^{-2} for the DC (direct current) component ($l = 0$) and 10^{-4} for the AC (alternating current) components ($l > 0$) of the spherical harmonics.

References

[1] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1

[2] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proc. arXiv:2410.11831*, 2024. 1

[3] 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. 1

[4] Diederik Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*. San Diego, California;, 2015. 2

[5] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1