Detail Enhanced Gaussian Splatting for Large-Scale Volumetric Capture

JULIEN PHILIP*, Eyeline Labs, United Kingdom
LI MA*, Eyeline Labs, United States of America
PASCAL CLAUSEN*, Eyeline Labs, Switzerland
WENQI XIAN*, Eyeline Labs, United States of America
AHMET LEVENT TAŞEL, Eyeline Labs, Canada
MINGMING HE, Eyeline Labs, United States of America
XUEMING YU, Eyeline Labs, United States of America
DAVID M. GEORGE, Eyeline Labs, United States of America
NING YU, Eyeline Labs, United States of America
OLIVER PILARSKI, Eyeline Labs, Germany
PAUL DEBEVEC, Eyeline Labs, United States of America

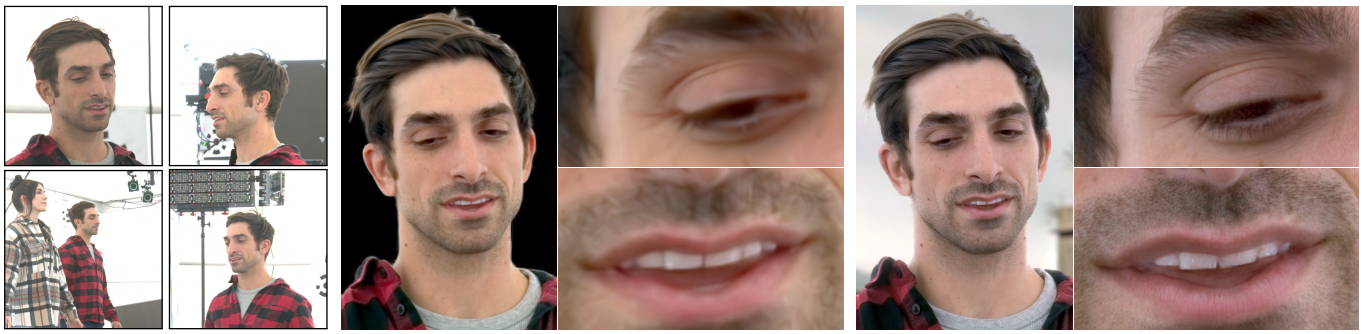

Fig. 1. From images taken in a dynamic multi-view capture stage (left), we reconstruct 4D Gaussian Splats tailored for production rendering (middle). The renderings are refined with a detail enhancement diffusion model before compositing (right). Insets show the extent of our detail enhancement. The enhancement model is trained on data of the actors acquired in a smaller facial capture stage.

We present a unique system for large-scale, multi-performer, high resolution 4D volumetric capture providing realistic free-viewpoint video up to and including 4K resolution facial closeups. To achieve this, we employ a novel volumetric capture, reconstruction and rendering pipeline based on Dynamic Gaussian Splatting and Diffusion-based Detail Enhancement. We design our pipeline specifically to meet the demands of high-end media production. We employ two capture rigs: the *Scene Rig*, which captures multi-actor performances at a resolution which falls short of 4K production quality, and the *Face Rig*, which records high-fidelity single-actor facial detail to serve as a reference for detail enhancement. We first reconstruct dynamic performances from the *Scene Rig* using 4D Gaussian Splatting, incorporating new model designs and training strategies to improve reconstruction, dynamic range, and rendering quality. Then to render high-quality images for facial closeups, we introduce a diffusion-based detail enhancement model. This model is fine-tuned with high-fidelity data from the same actors recorded in the *Face Rig*. We train on paired data generated from low- and high-quality Gaussian Splatting (GS) models, using the low-quality input to match the quality of the *Scene Rig*, with the high-quality GS as ground truth. Our results demonstrate the effectiveness of this pipeline in bridging the gap between the scalable performance capture of a large-scale rig and the high-resolution standards required for film and media production.

CCS Concepts: • **Computing methodologies** → **Rasterization**; **Image-based rendering**; **Image processing**; **Machine learning approaches**.

Additional Key Words and Phrases: Gaussian Splatting, Super Resolution

*These authors made equal technical contributions.

Authors' addresses: Julien Philip, Eyeline Labs, London, United Kingdom, julien.philip@scanlinevfx.com; Li Ma, Eyeline Labs, Los Angeles, United States of America, li.ma@scanlinevfx.com; Pascal Clausen, Eyeline Labs, Geneve, Switzerland, pascal.clausen@scanlinevfx.com; Wenqi Xian, Eyeline Labs, Los Angeles, United States of America, wenqi.xian@scanlinevfx.com; Ahmet Levent Taşel, Eyeline Labs, Vancouver, Canada, ahmet.tasel@scanlinevfx.com; Mingming He, Eyeline Labs, Los Angeles, United States of America, mingming.he@scanlinevfx.com; Xueming Yu, Eyeline Labs, Los Angeles, United States of America, xueming.yu@scanlinevfx.com; David M. George, Eyeline Labs, Los Angeles, United States of America, david.george@scanlinevfx.com; Ning Yu, Eyeline Labs, Los Angeles, United States of America, ning.yu@scanlinevfx.com; Oliver Pilarski, Eyeline Labs, Munich, Germany, oliver.pilarski@scanlinevfx.com; Paul Debevec, Eyeline Labs, Los Angeles, United States of America, debevec@scanlinevfx.com.

1 INTRODUCTION

4D volumetric performance capture systems are being leveraged with increasing frequency in media production, including for film and television where 4K resolution output is a requirement. Film and TV applications also introduce the need to capture the interaction of multiple actors over an extended area, and to produce recordings which appear high-resolution in wide, medium, and closeup shots. Placing the cameras around a larger performance area – ours is 6m × 9m – increases their distance to the subjects, which makes it harder to capture high-resolution details of the dynamic performances.

In this paper, we present a novel volumetric recording, reconstruction, and detail enhancement pipeline designed to address these challenges. Our approach leverages two complementary physical capture rigs built at different scales. The *Scene Rig* is designed for multi-view, multi-actor performance capture, enabling high-quality reconstuctions, but not enough to render production-quality facial closeups. The *Face Rig* records the head of each actor with production-quality resolution for closeups, but cannot capture full-body performances.

We first reconstruct performances of actors captured by the *Scene Rig* using a novel Gaussian Splatting-based approach optimized for this capture setup. This approach integrates a temporally stable camera calibration method and an HDR-aware 4D Gaussian Splatting method, which accounts for practical issues. Indeed, our rendering pipeline incorporates carefully designed components and training strategies optimized for color, exposure, and black levels, ensuring enhanced color fidelity that meets production needs.

Next, to bridge the quality gap between the *Scene Rig* captures and production resolution (especially for close-ups), we introduce a detail enhancement Diffusion Model. We modify a pre-trained image generation diffusion model, to support conditioning, to be temporally stable and to jointly generate RGB and Alpha channels. We fine-tune this model on high-fidelity *Face Rig* data of the actors who performed in the *Scene Rig*. Specifically, we use paired RGBA sequences of low-quality and high-quality renderings, obtained from pairs of low- and high-quality Dynamic Gaussian Splatting models. We limit the Gaussian count in the low-quality models to mimic the *Scene Rig*'s quality, with the high-quality renderings serving as ground truth. We demonstrate our method on several sequences of three sub-groups of actors showing various novel camera paths, including facial close-ups which significantly exceed the quality of the original *Scene Rig* capture. We demonstrate the importance of our system components through a set of baseline and ablation comparisons.

To summarize, the main contributions of this work are as follows:

- A two-stage approach to performance capture, combining a scene-scale capture rig and a single-actor facial capture rig.
- A novel high-quality scene-scale volumetric performance capture rig, incorporating both static and dynamic cameras to track the performance of multiple actors.
- A reconstruction pipeline for dynamic performance capture, featuring stable calibration of moving cameras and 4DGS with improved dynamic range and color fidelity.
- A detail enhancement Diffusion Model, which supports 4K, RGB and Alpha and with improved temporal stability.

2 RELATED WORK

2.1 Volumetric Data Capture and Reconstruction

Rendering photorealistic, view-controllable human performances from volumetric capture remains an active research area. Pioneering works focus on reconstructing 3D meshes from multi-view setups, addressing facial performance [Beeler et al. 2011; Fyffe et al. 2011; Guenter et al. 1998] and full-body geometry [Ahmed et al. 2008; Cagniart et al. 2010; de Aguiar et al. 2008; Kanade et al. 1997; Vlasic et al. 2008, 2009]. Some methods rely on template priors such as shape-from-silhouettes [Ahmed et al. 2008; Vlasic et al. 2008], or track and estimate the performer's deforming geometry using canonical or reference geometry [Beeler et al. 2011; Cagniart et al. 2010; de Aguiar et al. 2008; Vlasic et al. 2009]. Others leverage various illumination patterns to capture both geometry and reflectance information [Einarsson et al. 2006; Fyffe et al. 2011; Guo et al. 2019]. However, these approaches either rely on parameterized templates or fail to capture detailed geometry and appearance.

To reconstruct more details from multi-view videos, subsequent works propose using IR video cameras [Collet et al. 2015; Dou et al. 2017] or custom high-resolution depth sensors [Guo et al. 2019] to capture depth as geometry guidance, or incorporate custom color LED lights [Guo et al. 2019] Unfortunately, these mesh-based methods still struggle to reconstruct high-frequency details such as hair details, leading to a lack of photorealism.

With advances in neural rendering, learning-based approaches [Hedman et al. 2018; Kopanas et al. 2021; Lombardi et al. 2019; Rückert et al. 2022; Xu et al. 2019] have been proposed for novel view synthesis using multi-view capture, but are limited to static scenes.. To address dynamic objects, Lombardi et al. [2019] introduce a real-time capture and rendering system with a deep architecture. Recent methods [Meka et al. 2020; Zhao et al. 2022] combine traditional geometric pipelines with neural rendering, integrating volumetric and primitive-based rendering [Lombardi et al. 2021], or hybrid scene representations [Fridovich-Keil et al. 2023; Lin et al. 2023b; Peng et al. 2023; Xu et al. 2024a].

Among the new scene representations, neural radiance fields (NeRFs) [Mildenhall et al. 2020] have played a pivotal role in achieving high-quality novel view synthesis, with further improvements from InstantNGP[Müller et al. 2022], Mip-NeRF [Barron et al. 2021], Point-NeRF [Xu et al. 2022], Zip-NeRF [Barron et al. 2023], Deblur-NeRF [Ma et al. 2021], and RawNeRF [Mildenhall et al. 2022]. NeRFs are also used for performance synthesis [Du et al. 2021; Isik et al. 2023; Park et al. 2021a,b; Zhao et al. 2022]. Some approaches learn a canonical static template with deformable fields [Fang et al. 2022; Park et al. 2021a,b], showing promise for short videos but struggling with long sequences and complex motion. Others use 4D NeRFs [Du et al. 2021; Isik et al. 2023], incorporating timestamps with spatial location and view direction to model spatio-temporal changes.

Recently, 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023] emerged achieving a significant breakthrough in both rendering speed and quality. Further innovations followed, including Cinematic Gaussians [Wang et al. 2024b] that enhances the framework by introducing HDR rendering capabilities and depth awareness, and 4D Gaussian Splatting (4DGS) [Duan et al. 2024; Wu et al. 2024b] which extends to dynamic scenes by incorporating temporal consistency.

Recently, 4DGS has been improved, enhancing rendering quality [Li et al. 2024], reducing the number of input views [Mihajlovic et al. 2024], enabling streaming [Sun et al. 2024] and supporting longer sequences [He et al. 2024a; Shaw et al. 2024; Xu et al. 2024b]. These methods are constrained to synchronized static camera arrays and fail to meet key production requirements like linearity and 4K resolution. More aligned with our approach, Shen et al. [2024] uses diffusion-based video upsampling during training to enhance reconstruction quality, but is limited to 256×256 output. In contrast, we customize 4DGS with a new calibration strategy to handle dynamic cameras, HDR and capture artifacts. Then we apply detail enhancement as a post-process to add 4K details.

2.2 Image and Video Enhancement

Image and video enhancement can be achieved through super-resolution. Early video super-resolution approaches include recurrent-based [Haris et al. 2019; Huang et al. 2017; Liang et al. 2022; Sajjadi et al. 2018; Shi et al. 2022] and sliding-window-based [Caballero et al. 2017; Li et al. 2020; Liang et al. 2024; Xu et al. 2021; Yi et al. 2019] methods. Although both enhance their input, they tend to produce artifacts with real-world videos. RealBasicVSR [Chan et al. 2021, 2022] appends a pre-cleaning stage for artifact removal. RealViformer [Zhang and Yao 2024] uses covariance-based re-scaling for details. But they still face over-smoothing and temporal inconsistency. With the emergence of Diffusion models, recent efforts integrate diffusion priors with super-resolution [Lin et al. 2023a; Wang et al. 2024d; Wu et al. 2024a; Yang et al. 2023; Zhao et al. 2024]. StableSR [Wang et al. 2024d] couples a time-aware encoder and warping module with Stable Diffusion, while DiffBIR [Lin et al. 2023a] unifies generation and restoration via ControlNet. Some methods [Wu et al. 2024a; Yang et al. 2023] infuse semantic cues to refine details. Others [Chen et al. 2024; Yang et al. 2024a; Yuan et al. 2024; Zhou et al. 2024] augment text-to-image backbones [Ho et al. 2022; Rombach et al. 2022] with temporal layers or adapters, yet enforcing consistent details over long sequences remains difficult.

Text-to-video (T2V) priors also inspire video super-resolution. Xie et al. [2025] propose a spatial-temporal augmentation strategy, a local information enhancement module, and a dynamic frequency loss for artifact correction. Several works [He et al. 2024b; Wang et al. 2023] apply T2V to AI-generated videos. Upscale-A-Video [Zhou et al. 2024] relies on global latent propagation for longer sequences, while other methods [Chen et al. 2024; Yuan et al. 2024] embed U-Nets within a VAE pipeline. Despite gains in fidelity, artifact handling and long-range temporal smoothness remain open problems.

A line of work focuses on mitigating reconstruction artifacts in NeRFs or 3DGS by training restoration models. Zhou et al. [2023] introduce a dedicated degradation pipeline to simulate NeRF-specific artifacts. Catley-Chandar et al. [2024] pretrain on a small NeRF dataset and fine-tune per-scene to improve view consistency. Roessle et al. [2023] train a per-scene discriminator to enhance reconstructions. However, these models primarily refine input signals and can't generate new details beyond the input images which we require. Wu et al. [2025] trains a diffusion model on large-scale 3D reconstructions for generalizable artifact removal. Because of the generality of this approach, the added details are not based on specific subject appearance, but on general priors.

For production-quality high-resolution generation, tuning-based methods [Guo et al. 2024; Hoogeboom et al. 2023; Liu et al. 2024; Ren et al. 2024; Teng et al. 2023; Zheng et al. 2024] fine-tune large models on scarce high-resolution data, while tuning-free pipelines [Cao et al. 2024; Du et al. 2024; Haji-Ali et al. 2024; He et al. 2024c; Hwang et al. 2024; Jin et al. 2023; Lee et al. 2023; Lin et al. 2024] perform patch-wise inference. Some methods mitigate block artifacts [Bar-Tal et al. 2023; Du et al. 2024], expand receptive fields [He et al. 2024c], and manipulate latents across scales [Huang et al. 2024]. Pivoting strategies [Guo et al. 2025; Qiu et al. 2024] fuse stable low-resolution semantics with multi-scale upsamplers, but balancing local detail with global consistency remains a challenge in super-resolution. Moreover, to the best of our knowledge, no video-based super-resolution method supports 4K. Instead, we modify an image-based model [Labs 2024] with a frame warping scheme and low frequency swapping for temporal stability.

3 OVERVIEW

The workflow of our method, from volumetric data capture to dynamic scene reconstruction and enhancement, is illustrated in Fig. 2. We use both a *Scene Rig* for multi-actor dynamic scenes and a *Face Rig* to capturing high-resolution facial details. First, the *Scene Rig* captures a multi-view performance, allowing actors to move freely. We adopt a variation of the 4DGS method, *Poly4DGS*, to reconstruct the dynamic scene with globally and temporally consistent structure but limited facial details.

Next, the *Face Rig* captures high-resolution facial data of the same actors within a small volume. This data is used to train a diffusion-based detail enhancement model, which we use to enhance the *Scene Rig* renderings.

4 DYNAMIC RECONSTRUCTION AND RENDERING

Our first stage aims at reconstructing and rendering actors' free-form performances using 4DGS. The quality and temporal consistency of 4DGS is highly correlated with the quality of the input data. Unfortunately, capturing fast-moving actors in a multi-view setup is challenging due to motion blur, defocus, and actors moving out of frame. To tackle these challenges, we use a *Scene Rig* combining static and dynamically aimed cameras to record the actors in the highest resolution possible. Given the acquired sequences, we introduce a new camera calibration method that leverages a clean background to ensure accurate and consistent camera parameters. We then reconstruct the performance using *Poly4DGS* allowing to render dynamic scenes from arbitrary viewpoints.

4.1 Hardware Setup of the *Scene Rig*

The *Scene Rig* features 180 synchronized 4K Z-CAM e2 cinema cameras 30 of these cameras are mounted on the ceiling and 25 cameras are placed on the floor, providing top-down and bottom-up perspectives to minimize occlusions.

The remaining cameras are mounted on 12 mobile carts arranged circularly around the $6m \times 8m$ stage, allowing our system to adapt

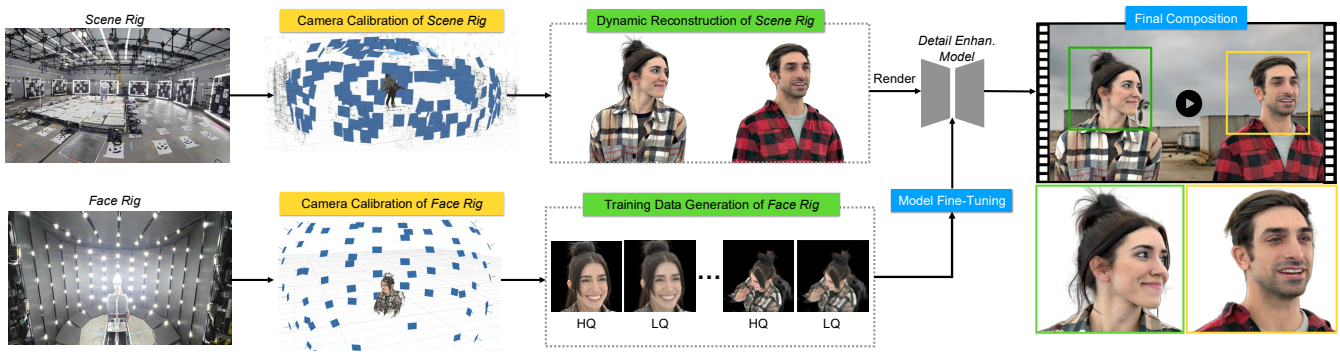

Fig. 2. Pipeline Overview. Actors perform in the *Scene Rig*, where full-body performances are captured. Using our Poly4DGS framework, we reconstruct the **Performance**. The same actors are then captured in the *Face Rig*. We generate Poly4DGS models for a portion of their facial performance: a high-quality model (**HQGS**, 4M Gaussians) and a low-quality model (**LQGS**, 50K-200K Gaussians). These reconstructions are used to train an **Image Enhancement Module** which refines the renderings of the low-quality GS to be like the high quality one. Finally, the trained model is used to enhance renderings from the 4DGS performance. Please refer to our supplementary video for the final composition of 4K render results.

to different configurations. The cameras on each cart follow a structured zig-zag pattern, with slight vertical offsets between adjacent units. Calibration patterns are placed throughout the room and stage to ease camera alignment.

Our stage features two types of cameras: static wide field-of-view (FoV) cameras and tracking cameras with instrumented pan, tilt, zoom, and focus. The 90 static cameras are equipped with 14-42mm lenses, providing broad coverage of the entire stage. Additionally, 40 landscape tracking cameras with 45-175mm zoom lenses dynamically adjust the camera pan, tilt, zoom, and focus to track the actors' faces. Tracking is performed using unobtrusive OptiTrack™ markers placed on the clothing below the back of the neck of actors. Finally, 50 additional portrait orientation cameras track and frame full-body, upper-body, and lower-body movements.

The stage is uniformly lit by white LED strips mounted on the carts and ceiling, ensuring consistent illumination across the stage. The LED strips emit 1 ms flashes in sync with the 24 fps camera shutters to reduce motion blur, strobing at 72 Hz to exceed the flicker fusion frequency for actor comfort.

4.2 Volumetric Data Capture and Processing

We direct three groups of actors to perform diverse actions, including running, walking, talking, and playing basketball. The footage is recorded at 4K, 24 fps. We also capture a color chart to calibrate color between the *Scene Rig* and the *Face Rig*.

Camera alignment and calibration. Camera alignment of our volumetric capture system is challenging due to dynamic cameras with varying zoom levels. The shallow depth of field in zoomed-in views blurs the background, making these views difficult to calibrate. Additionally, focal length changes can be confused with camera movement, leading to ambiguities between intrinsics and extrinsics. Moving cameras further complicate the process, requiring frame-by-frame re-calibration. Without proper constraints, this can lead to alignment inconsistencies, such as drift and stutter.

To align both static and dynamic cameras, we use a two-step process. First, we calibrate only the static wide-FoV cameras, using lidar scans and the middle frame of the sequence while keeping their focal lengths fixed. Once calibrated, we fix the static cameras' position and then calibrate the tracking cameras separately. We ensure that zoom changes are properly accounted for rather than freely estimated per frame. To do so, we fit a smooth function to the changes in focal length provided by the cameras and use it as a regularization across frames, preventing sudden jumps or inconsistencies. We find that by doing so, the number of identified points in the initial point cloud increases by nearly 50%.

4.3 Dynamic Scene Reconstruction Using *Poly4DGS*

Given multi-view video data as input, we design the dynamic scene reconstruction method, *Poly4DGS*, based on advanced 3DGS [Kerbl et al. 2023] and 4DGS [Duan et al. 2024; Yang et al. 2024c] techniques. We also incorporate new training strategies, including color space adjustment, exposure, and black-level optimization, to maximize reconstruction quality while preserving temporal consistency.

4.3.1 Poly4DGS.

We represent dynamic 3D scenes using a set of 4D primitives. Unlike previous approaches that parameterize each 4D primitive as Gaussian functions in 4D space using 4D-Rotor [Duan et al. 2024] or dual-quaternion [Yang et al. 2024c], we adopt a simplified representation that is easy to implement based on existing 3DGS rasterizer [Ye et al. 2025]. Our method directly parametrizes the motion of each Gaussian as polynomial expansions over time for each property. Formally, we define:

$$\mu_t = \Sigma_{i=0}^{n_\mu} \mu_i (t - t_0)^i, \quad \mathbf{q}_t = \Sigma_{i=0}^{n_\mathbf{q}} \mathbf{q}_i (t - t_0)^i,$$
$$\mathbf{s}_t = \Sigma_{i=0}^{n_\mathbf{s}} \mathbf{s}_i (t - t_0)^i, \quad \mathbf{o}_t = \mathbf{o}_0 e^{-\frac{1}{2}\Sigma_{i=1}^{n_\mathbf{o}} \lambda_i (t - t_0)^{2i}},$$

where μ_t, $\mathbf{q}_t$, $\mathbf{s}_t$, and $\mathbf{o}_t$ represent the mean, quaternion rotation, scaling, and opacity of a 3D Gaussian at timestamp t, respectively. Intuitively, the parameter t_0 denotes the temporal center of the Gaussian, and λ_i controls the length of the Gaussian lifespan. Notably, Duan et al. [2024] have shown that when $n_\mu = 1, n_\mathbf{q} = 0, n_\mathbf{s} = 0$ and $n_\mathbf{o} = 1$, this parameterization corresponds to slicing a 4D Gaussian at time t. By introducing higher-order terms for different properties, our approach allows more flexibility, thus enabling each primitive

to capture more complex motions. Empirically, we find $n_\mu = 2$, $n_\mathbf{q} = 1$, $n_\mathbf{s} = 0$, and $n_\mathbf{o} = 2$ result in a slight improvement in PSNR (+0.2dB) without introducing visible overhead. Since performance sequences exhibit minimal texture changes, we use time-constant spherical harmonics. We use the standard Gaussian splatting rendering [Kerbl et al. 2023] with antialiasing [Yu et al. 2024] to rasterize the 3D Gaussians at time t and we use Kheradmand et al. [2024] to handle pruning and relocation.

With *Poly4DGS*, the number of primitives required to represent a 4D sequence with comparable quality scales approximately linearly with the length of the sequence. To balance representation quality and computational feasibility, we divide the whole sequence into smaller segments and train each segment independently. The segmentation ensures high-quality results but also enhances the scalability, as the training of segments can be parallelized efficiently on GPU clusters.

We initialize 4D primitives from a sequence of dense point clouds reconstructed per frame. We observe that using dense point clouds of approximately 200k points leads to much faster convergence compared to sparse point clouds.

4.3.2 *Training Color Space.*

We aim to train our 4DGS with linear OpenEXR files to maintain the input's dynamic range while ensuring proper control over the training color space. The direct solution of training Gaussian Splatting in linear color space results in poorer outcomes, typically we observed a 5dB decrease in PSNR as detailed in supplemental materials. We hypothesize this is due to a conditioning issue in the optimization process, similar in spirit to findings for camera optimization and *floaters* [Park 2023; Philip and Deschaintre 2023]. To address this, we propose storing color parameters in an unbounded tone-mapped space, but rasterizing using linearized colors. We experimented with several color spaces such as $log(1+x)$ and *sRGB* without clamping, and found they all performed equally. To maintain a single tone-mapping and inverse operator we used the unbounded *sRGB* color space. Specifically, the Gaussian colors, c_G, obtained after evaluating the spherical harmonics, are transformed to linear space using the inverse sRGB mapping f_srgb^{-1}, before being rasterized:

$$c_\mathrm{linear} = f_\mathrm{srgb}^{-1}(c_\mathrm{G}) \quad \text{and} \quad C_\mathrm{rast} = \mathcal{R}(c_\mathrm{linear}, \ldots), \tag{1}$$

where $\mathcal{R}$ represents the rasterization. f_srgb^{-1} operates in $\mathbb{R}^+$ without clamping. We leave out other Gaussian attributes for simplicity.

Camera 1 Camera 2

Fig. 3. The *Scene Rig* captures suffer from severe lens glare for some cameras.

4.3.3 *Exposure and Black-Level Optimization.*

The zoom lens optics produce a noticeable amount of veiling glare and lens flares in our input frames, as visible in Fig. 3. We propose to correct for this using 32×32 spatially-varying exposure and black-level adjustments grids, also accounting for sensor variations:

$$C_\mathrm{adjusted} = (C_\mathrm{rast} + B^c(x,y)) \cdot E^c(x,y), \tag{2}$$

where $B^c(x,y)$ and $E^c(x,y)$ indicate the black level and exposure at pixel position (x,y), and $\cdot$ is the Hadamard product. We upsample the grid, by applying a FFT to it, zero-padding to C_rast size and applying the IFFT. The backward pass of this process is orders of magnitude faster than the slicing operation used by bilateral grids [Wang et al. 2024a] at high resolution.

The adjusted color can then be transformed to a display color such as sRGB; combining all the steps, the final output color C_final is:

$$C_\mathrm{final} = f_\mathrm{srgb}(C_\mathrm{adjusted}) = f_\mathrm{srgb}\left(\left(\mathcal{R}\left(f_\mathrm{srgb}^{-1}(c_G)\right) + B^c\right) \cdot E^c\right). \tag{3}$$

More details about the tone-mapping, and the impact of the GS storage color space are provided in the supplemental document

4.3.4 *Training Objective.*

Our training objective builds upon the standard 3DGS reconstruction loss, $\mathcal{L}_\mathrm{recon}$, which combines L_1 and L_SSIM. The reconstruction loss $\mathcal{L}_\mathrm{recon}$ is computed in unbounded sRGB space. Additionally, we include the following regularization terms:

Exposure and black-level regularization: We encourage the exposure maps $E^c(x,y)$ to have an average value of 1 across all spatial and camera dimensions, to ensure stability. We also encourage higher black-level offsets $B^c(x,y)$ for the current frame, to reduce glare:

$$\mathcal{L}_\mathrm{exposure} = \left(\sum_{x,y,c} \frac{E^c(x,y) - 1}{N_x N_y N_c}\right)^2 \qquad \mathcal{L}_\mathrm{black} = -\frac{1}{N}\sum_{pixels} B^c(x,y) \tag{4}$$

Our final training objective is a weighted sum of all components:

$$\mathcal{L}_\mathrm{total} = \mathcal{L}_\mathrm{recon} + \lambda_e \mathcal{L}_\mathrm{exposure} + \lambda_b \mathcal{L}_\mathrm{black} \tag{5}$$

where $\lambda_e = 10$, and $\lambda_b = 0.05$ are hyperparameters controlling the contributions of each regularization term.

5 DETAIL ENHANCEMENT

Although we aim to maximize the quality of novel-view synthesis in the first stage, the level of detail achieved remains limited by the large scale of the capture area, the finite number of cameras, and imperfections in lens optics. To obtain production-level details, particularly for facial close-ups, we use a smaller-scale multi-view capture stage, the *Face Rig*. From the *Face Rig* data, we generate a paired training dataset of sequences using low- and high-quality Dynamic GS reconstructions. This dataset is used to train a Detail Enhancement model finetuned from an Image Diffusion model. Its role is to remove Gaussian artifacts from lower-quality *Scene Rig* renderings and generate fine details, enhancing visual photorealism.

To improve the model's temporal stability, we condition it on an Optical Flow reprojection of the preceding generated frame. Finally, to help compositing, we modify the model architecture to jointly predict the detailed image and its corresponding alpha.

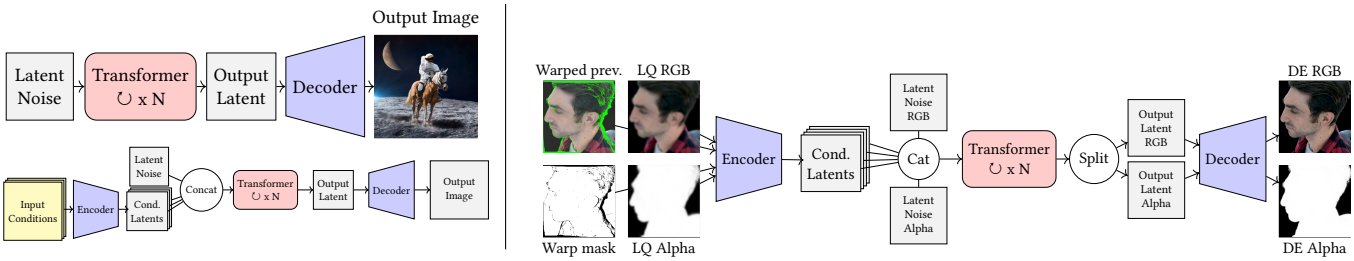

Fig. 4. Architectural changes made to the base Flux model [Labs 2024]. Starting from the Latent Diffusion Architecture (top left), we add input channels to condition the network (bottom left). To improve temporal stability and generate an alpha channel, we condition our model on the previous warped output, a validity mask, and both the LQ RGB and Alpha. We also double the size of the latent space to predict RGB and Alpha jointly (right).

Fig. 5. Illustration of the *Face Rig* and corresponding reconstructions. From left to right: the *Face Rig* hardware, an input image, a low-quality GS render and the corresponding high-quality GS render used for supervision.

5.1 Hardware Setup of the *Face Rig*

The *Face Rig* is a volumetric capture stage with a multi-view camera array placed within a capped cylinder of black LED panels and white LED lights as shown in Fig. 5, left. The 80 LED panels (16 columns, each consisting of 5 panels) form a capture volume approximately 250cm tall and 276cm in diameter, with the ceiling and floor covered with 30 and 10 panels, respectively. This stage features 75 synchronized 4K Z-CAM E2 cinema cameras, evenly installed and peering through 5cm gaps in the walls and ceiling. To ensure consistent flat lighting for facial data capture, 110 white LED lights are evenly distributed throughout the stage.

5.2 Volumetric Data Capture and Processing

After performing in the *Scene Rig*, we ask the actors to individually perform a short sequence of diverse facial expressions in the *Face Rig* . This footage is captured at 24fps 4K resolution. We also capture a color chart for color calibration against the *Scene Rig*.

From each actor's performance, we uniformly sample 60 subsequences of 8 frames, ensuring a diverse range of expressions. For each selected sub-sequence, we reconstruct two *Poly4DGS* models: a high-quality model with 1 million Gaussians per-frame and a lower-quality model with a number of Gaussians sampled from 50,000 to 200,000. This range corresponds to the minimum and maximum number of Gaussians left when cropping the upper-body of actors in the *Scene Rig* reconstructions. This ensures that the low quality reconstructions match the quality expected at test time.

5.3 Training Data Generation

To obtain training data for our detail enhancement model, we render paired sequences of low- and high-quality GS in a 16-bit RGBA EXR format at 4096×4096 resolution. For each reconstructed subsequence, we generate 60 camera paths each spanning 8 frames, with a moving camera focusing on the face. In total, this results in approximately 3600 paired sequences of low- and high-quality renderings per actor, covering a diverse range of Gaussians artifacts, facial expressions, camera motions, and focal lengths. After rendering, the dataset contains approximately 3600 paired sequences of low- and high-quality renderings per actor. Additional details about the camera paths and data generation are included in the supplemental material.

5.4 Detail Enhancement Model

The goal of our detail enhancement model is to remove the Gaussian-like artifacts from the *Scene Rig* renderings and restore fine details, especially useful for rendering close-up shots.

One promising approach for this task would be to use a conditioned video diffusion model [Blattmann et al. 2023; Jin et al. 2024; Yang et al. 2024b]. To meet the production quality requirements, we aim to enhance our images at 4K resolution. However, video diffusion models are memory-intensive and resolution-limited, typically ranging from 480p to 768p, well below our target of 2160p.

We thus leverage an image diffusion model, Flux [Labs 2024], which natively supports resolutions from 1080p to 1440p and incorporates strong natural image priors. We introduce key architectural changes to this pretrained model to address the main challenges we face. First, we need to condition the model on the renderings from the *Scene Rig*. Second, we aim to obtain a temporally stable output, as generating frames independently causes temporal flickering. Finally, as the model will be used for virtual production, we seek to enhance the Alpha channel together with RGB to ease the composition with backgrounds.

5.4.1 Image-Conditioning. As shown in the inference generation scheme of Flux [Labs 2024] in Fig. 4 (top left), latent noise is sampled from a normal distribution and then denoised multiple times by the diffusion transformer, to produce an output latent, which is converted to an image using the VAE Decoder. Following the existing works [Luo et al. 2024; Zeng et al. 2024], we modify the Diffusion Transformers (DiT) to condition it on multiple images. We extend the weight matrix of the first linear layer so it takes as input, $(N +$

1) $\mathcal{L}_{ch}$ channels instead of $\mathcal{L}_{ch}$, where N represents the number of conditioning images added.

The inference scheme for this model is shown in Fig. 4 (bottom left). The input conditions are encoded separately with the VAE and concatenated to the latent noise, before being processed.

5.4.2 Temporal Stability. We first explore training a model conditioned on low-quality GS to predict high-quality renderings. While this yields high-quality results for single frames, as shown in our supplemental video, the output lacks temporal stability.

Optical Flow Warping. To make the output of our model temporally consistent, especially for fine details, we propose conditioning it on the previous output frame, if available. Therefore, we add two additional conditions to the model: a warped version of the previous output frame and a validity mask. We compute the optical flow [Teed and Deng 2020] between the current and previous input low-quality (LQ) renderings. We then resample both the previous LQ and enhanced renderings. The resampled LQ rendering is compared to the current one to generate the validity mask. For the first frame, both conditions are replaced with zeros, and the network is trained with 50% dropout on these inputs.

Low-Frequency Stabilization. Even with this guidance, the output suffers from low-frequency flicker, such as global intensity shifts. As we show in the supplemental material, this is partially caused by the VAE's inability to preserve low-frequency information through encoding and decoding. To alleviate this, we compute a 5-level Laplacian pyramid for both the LQ input and detail-enhanced result and then swap the lowest level (120×68 pixels) of the output with that of the input. Experiments show this simple approach is highly effective in terms of improving temporal stability.

5.4.3 Alpha Channel Enhancement. Since we aim to composite the final frames onto virtual backgrounds, an accurate Alpha channel is essential. To enhance the Alpha channel of the LQ input, we modify the model to additionally condition on this channel. We also double the number of input latent channels, similar to adding an extra input condition. Finally, we adjust the output linear layer to predict twice the number of channels. Our final architecture, shown in Fig. 4 (right), takes four input conditions - LQ RGB, LQ Alpha, warped previous output, and warped validity mask - along with two latent noise inputs, resulting in $6 \times \mathcal{L}_{ch}$ input channels. It outputs two latents, *i.e.*, $2 \times \mathcal{L}_{ch}$ channels. The Alpha and RGB channels are decoded independently by the decoder.

The model is trained simultaneously on all seven actors, with a different text prompt for each subgroup. Additional implementation details are provided in the supplemental materials.

6 EXPERIMENTS

6.1 Main Results

We present the results of eight different performances from three groups of two or three actors. For each performance, we design virtual camera paths using a custom Maya [Autodesk, INC. 2024] plugin. We then render our reconstructions of the performances – cropped to delete the background – and enhance the renderings with our detail enhancement model. In Fig. 6 (best viewed full screen) we present qualitative image results. We show our reconstruction quality matches the capture resolution while providing better colors. The detail enhancement model effectively removes GS artifacts and adds significant missing detail both in the RGB and Alpha layers. The results are then composited by an artist onto a background sequence using the same camera path. For backgrounds, we used 3DGS reconstructions or generated videos [Runway AI, Inc. 2024]. Additional results are presented in the accompanying video. We provide implementation details and estimates for the compute overhead of our method in the supplemental.

6.2 Baselines/Ablations and comparisons

We conduct comparative and ablative analysis across different design choices in our method to evaluate the performance of the dynamic scene reconstruction and detail enhancement models.

Table 1. Ablation results of *Poly4DGS* reconstruction. The best results are marked in **bold**. All our added components contribute to the final quality of the reconstruction. Using a per-frame GS instead of 4DGS leads to poorer reconstruction.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Full	**27.29**	**.9752**	**.0160**
w/o black level	24.11	.9576	.0200
w/o expo. opt.	25.50	.9718	.0161
w/o dense pt.	25.31	.9576	.0288
per-frame GS	26.11	.9697	.0204
4D-Rotor [Duan et al. 2024]	27.08	.9730	.0171

6.2.1 Dynamic Scene Reconstruction. To evaluate our reconstruction pipeline, and main additions we conduct an ablation study. We hold out 12 out of 140 cameras, covering both close-up and large-scale FoV. Since our optimized GS model has a different exposure and black level with respect to the input images, we optimize the exposure and black level of the held-out views before comparing to ground truth. We report the *PSNR*, *SSIM* and *LPIPS* averaged over three sequences in Tab. 1 for different variants of our pipeline. Without black-level optimization (w/o black level) and exposure optimization (w/o expo. opt.) the view-inconsistent effects such as lens glare and exposure variations get baked into the splats, leading to low contrast and poor colors, also shown in Fig. 7. Initializing from dense point cloud (w/o dense pt.) helps to reconstruct fine details. Last we compare with using per-frame 3DGS [Kerbl et al. 2023] reconstruction and 4D-Rotor [Duan et al. 2024] both modified to benefit from dense point cloud initialization and black level and exposure optimization. We observe that using the *Poly4DGS* formulation achieves better performance. This is because 4DGS allows Gaussian primitives to be shared across frames with a smoother interpolation, leading to a higher per-frame gaussian count. Most importantly, as shown in the supplementary video, our 4DGS produces more temporally stable results compared to per-frame 3DGS.

6.2.2 Detail enhancement. We compare our detail enhancement model to existing SoTA super-resolution methods, which can be categorized in image-based [Wang et al. 2024c; Yue et al. 2024] and video-based categories [Feng et al. 2024; Zhou et al. 2024]. We also

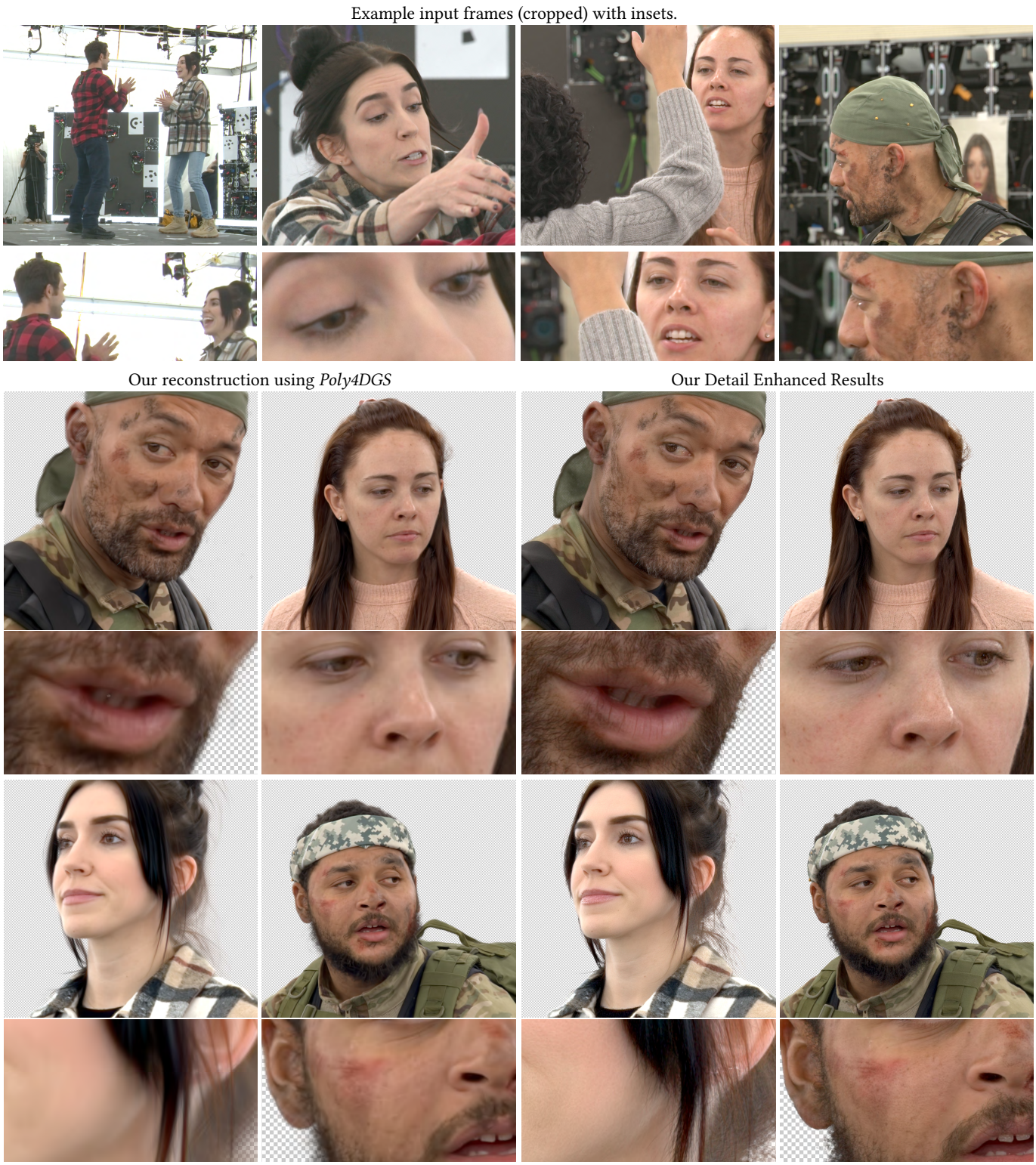

Fig. 6. Top: input training views captured in our *Scene Rig*. Bottom-left: our *Poly4DGS* reconstruction with insets. Bottom-right: final results using our super-resolution module and compositing. We can observe that the Gaussian artifacts present at extreme zoom levels are effectively removed and new details added.

Fig. 7. Qualitative results for several ablation studies. Our full method achieves the highest quality, delivering good color contrast and sharp details. Without initializing from dense point cloud (w/o dense pt.), the method fails to reconstruct fine details like the stripe on the girl's shirt. Without black-level and exposure optimization (w/o black level and w/o expo. opt.), view inconsistencies caused by camera exposure variations and lens glare are baked in the GS.

Table 2. Ablations and Comparison of the detail enhancement model using temporal and image quality metrics. MS, TF, MUSIQ and NIQE were copmputed on the *Scene Rig* data in a No-Reference manner, while PSNR, SSIM, LPIPS, TPSNR, FID and FVD were computed on the test of the *Face Rig* data. Best values are in **bold** and second best underlined.

Method	**No-Ref metrics using *Face Rig* data**				**Ref-based metrics using *Scene Rig* data**					
	MS ↑	TF ↑	MUSIQ ↑	NIQE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	TPSNR ↑	FID ↓	FVD ↓
Ours	**95.34**	**94.29**	51.79	**5.28**	31.29	0.8871	0.1812	31.27	102.77	213.52
w/o OFW	94.72	94.06	52.24	5.48	**31.43**	**0.8872**	0.1614	**31.42**	103.44	**202.90**
w/o OFW, w/o LFS	94.73	93.90	**53.17**	5.89	30.61	0.8604	**0.1612**	30.60	**99.88**	209.85
Ours	**95.34**	**94.29**	51.79	**5.28**	31.29	0.8871	**0.1812**	31.27	102.77	**213.52**
ResShift 2024	93.70	92.85	58.83	5.82	29.33	0.7771	0.3221	29.33	200.16	662.70
ResShift (FT) 2024	93.36	92.34	59.12	5.77	29.78	0.8517	0.2146	29.78	**87.93**	405.16
SinSR 2024c	93.89	93.08	**62.59**	5.32	29.19	0.7070	0.3521	29.19	199.17	656.69
Upscale-A-Video 2024	94.64	93.66	34.87	7.87	30.30	0.8329	0.3620	30.29	191.96	341.74
KEEP 2024	94.09	93.20	45.33	6.16	30.08	0.8627	0.2996	30.07	176.90	350.91
Input Poly4DGS Sequence	94.18	93.25	37.22	6.89	**31.72**	**0.9009**	0.2474	**31.72**	150.51	333.11

compare our approach to a version of ResShift [Yue et al. 2024] fine-tuned (FT) on 512-resolution crops of our dataset, which was trained from scratch for approximately 520K steps, nearing convergence. Additionally, we provide 2 ablated versions of our method: 1.) removing the Optical Flow Warping mechanism (w/o OFW) 2.) removing the Optical Flow, and the Low Frequency Swapping (LFS) (w/o OFW, w/o LFS), which corresponds to an image model conditioned on the low quality inputs.

We show quantitative results in Tab. 2. We quantitatively evaluate our method using both No-Reference metrics and Reference based Metrics. The no-reference scenario allows us to evaluate our model in its natural usage setting using renderings of the *Scene Rig* reconstruction as input. We used all our render paths for evaluation. We first provide temporal stability metrics, namely motion smoothness (MS) and temporal flickering (TF) from the widely adopted VBench [Huang et al. 2023] benchmark. We also provide two no-reference image quality metrics, MUSIQ [Huang et al. 2023] and NIQE [Mittal et al. 2013].

Referenced based metrics are computed on *Face Rig* test data, left out of the training, and allows us to evaluate the PSNR, SSIM, LPIPS [Zhang et al. 2018], TPSNR [Hasselgren et al. 2020], FID, and FVD [Unterthiner et al. 2019] metrics. PSNR, SSIM, LPIPS, and temporal PSNR (which is PSNR computed on the temporal finite differences) were calculated per frame and averaged over all test frames and sequences. For both FID and FVD we extract features from patches and compute the distance over all patches.

From the ablations (Tab. 2, first three lines) we can see that our method does better at temporal consistency on the MS and TF metrics at the cost of a slight decrease in image quality compared to removing the Optical Flow Warping, this is due to a slight image smoothing. Swapping the Low Frequency without Optical Flow Warping strongly helps temporal consistency, though the Warping further improves it.

Our method greatly outperforms all other methods and baselines in terms of reference-based image quality and is competitive with Upscale-A-Video in temporal stability, though this method produces much less details results as can be seen in our supplemental video. Our method is also the best in the NIQE metric. ResShift and SinSR beat us on MUSIQ, which we attribute to the overly smooth and sharp nature of SinSR and ResShift; they exhibit less details and are

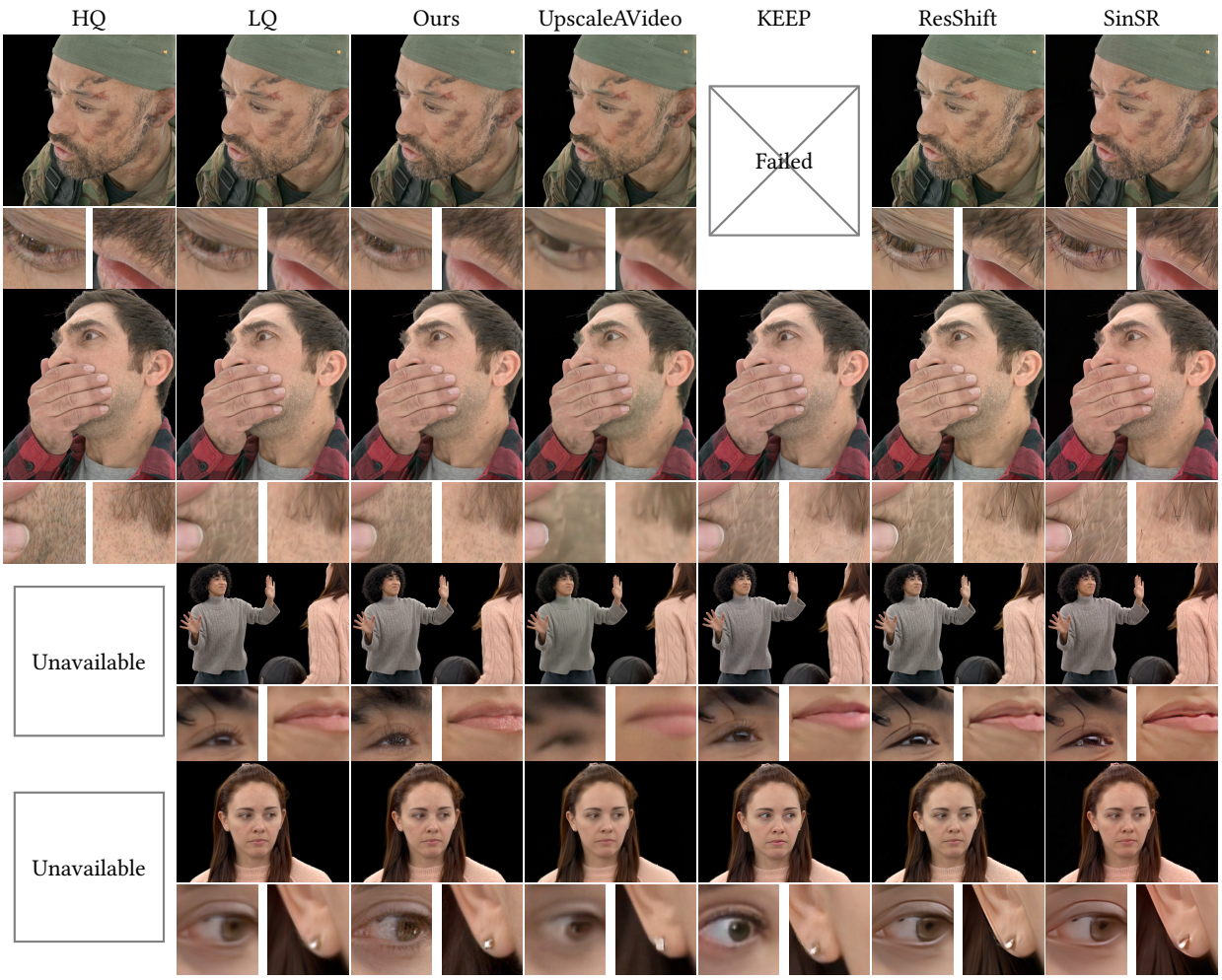

Fig. 8. Comparisons for our Detail Enhancement module. Top two images are from the validation set, while bottom two are from a test sequence. From left to right: high resolution GS render, typically used as ground truth during training. A low res GS render used as input for the following. Our result. UpscaleAVideo [Zhou et al. 2024]. Results from KEEP [Feng et al. 2024]. Results using a ResShift[Yue et al. 2024] model. And SinSR results [Wang et al. 2024c].

less realistic as seen in our supplemental video. We are significantly better than the input Poly4DGS on these metrics though. The input Poly4DGS without detail enhancement beats our result in terms of PSNR and SSIM, but this is expected given it is mostly a "blurred" version of the GT target, while our method generates details, such as hair strand, that might not be aligned with the ground truth. Our method has a better LPIPS metric, which captures the visual similarity to the ground truth.

We present qualitative comparisons in our supplemental video and Fig.8. We can see our method is temporally stable and produces higher quality, detailed images compared to other methods.

7 CONCLUSION, LIMITATIONS, AND FUTURE WORK

We have presented a novel 4D performance capture pipeline that bridges the gap between scalable dynamic capture and production-quality rendering using a detail enhancement model. We demonstrate added fine details thanks to our enhancement model, allowing us render frames that match production requirements. Nonetheless, our method has some limitations. It is hardware and resource-intensive, making it not universally accessible, but still appropriate for use in professional production. While our temporal stability mechanisms greatly help, some flicker may remain on silhouette. We did not explore relighting in this work, leaving it to future research (e.g. [He et al. 2024a]), which will allow for better integration into various backgrounds. Eye reflections showing the sparse lighting of the *Face Rig* are transferred with detail enhancement, suggesting that future work should provide a way to synthesize realistic eye reflections for a given environment. A specialized model could be used [Li et al. 2022] and adding continuous lighting patterns from LED panels could mitigate the sparsity of the reflections.

REFERENCES

Naveed Ahmed, Christian Theobalt, Christian Rössl, Sebastian Thrun, and Hans-Peter Seidel. 2008. Dense correspondence finding for parametrization-free animation reconstruction from video. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society. https://doi.org/10.1109/CVPR.2008.4587758

Autodesk, INC. 2024. *Maya*. https://autodesk.com/maya

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *arXiv preprint arXiv:2302.08113* (2023).

Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. *ICCV* (2021).

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2023. Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. *ICCV* (2023).

Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul A. Beardsley, Craig Gotsman, Robert W. Sumner, and Markus H. Gross. 2011. High-quality passive facial performance capture using anchor frames. *ACM Trans. Graph.* 30, 4 (2011), 75. https://doi.org/10.1145/2010324.1964970

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*. 4778–4787.

Cedric Cagniart, Edmond Boyer, and Slobodan Ilic. 2010. Probabilistic Deformable Surface Tracking from Multiple Videos. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV (Lecture Notes in Computer Science, Vol. 6314)*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.). Springer, 326–339. https://doi.org/10.1007/978-3-642-15561-1_24

Boyuan Cao, Jiaxin Ye, Yujie Wei, and Hongming Shan. 2024. AP-LDM: Attentive and Progressive Latent Diffusion Model for Training-Free High-Resolution Image Generation. *arXiv preprint arXiv:2410.06055* (2024).

Sibi Catley-Chandar, Richard Shaw, Gregory Slabaugh, and Eduardo Pérez-Pellitero. 2024. RoGUENeRF: A Robust Geometry-Consistent Universal Enhancer for NeRF. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XII* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 54–71. https://doi.org/10.1007/978-3-031-73254-6_4

Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2021. Basicvsr: The search for essential components in video super-resolution and beyond. In *CVPR*. 4947–4956.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. 2022. Investigating tradeoffs in real-world video super-resolution. In *CVPR*. 5962–5971.

Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. 2024. Learning Spatial Adaptation and Temporal Coherence in Diffusion Models for Video Super-Resolution. In *CVPR*. 9232–9241.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam G. Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM Trans. Graph.* 34, 4 (2015), 69:1–69:13. https://doi.org/10.1145/2766945

Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance capture from sparse multi-view video. *ACM Trans. Graph.* 27, 3 (2008), 98. https://doi.org/10.1145/1360612.1360697

Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. 2017. Motion2fusion: real-time volumetric performance capture. *ACM Trans. Graph.* 36, 6 (2017), 246:1–246:16. https://doi.org/10.1145/3130800.3130801

Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. 2024. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6159–6168.

Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. 2021. Neural Radiance Flow for 4D View Synthesis and Video Processing. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 14304–14314. https://doi.org/10.1109/ICCV48922.2021.01406

Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 2024. 4D-Rotor Gaussian Splatting: Towards Efficient Novel View Synthesis for Dynamic Scenes. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 87, 11 pages. https://doi.org/10.1145/3641519.3657463

Per Einarsson, Charles-Felix Chabert, Andrew Jones, Wan-Chun Ma, Bruce Lamond, Tim Hawkins, Mark Bolas, Sebastian Sylwan, and Paul Debevec. 2006. Relighting human locomotion with flowed reflectance fields. In *Proceedings of the 17th Eurographics Conference on Rendering Techniques* (Nicosia, Cyprus) *(EGSR '06)*. Eurographics Association, Goslar, DEU, 183–194.

Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. 2022. Fast Dynamic Radiance Fields with Time-Aware Neural Voxels. In *SIGGRAPH Asia 2022 Conference Papers*.

Ruicheng Feng, Chongyi Li, and Chen Change Loy. 2024. Kalman-Inspired Feature Propagation for Video Face Super-Resolution. arXiv:2408.05205 [cs.CV] https://arxiv.org/abs/2408.05205

Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. 2023. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 12479–12488. https://doi.org/10.1109/CVPR52729.2023.01201

Graham Fyffe, Tim Hawkins, Chris Watts, Wan-Chun Ma, and Paul E. Debevec. 2011. Comprehensive Facial Performance Capture. *Comput. Graph. Forum* 30, 2 (2011), 425–434. https://doi.org/10.1111/J.1467-8659.2011.01888.X

Brian Guenter, Cindy Grimm, Daniel Wood, Henrique Malvar, and Fredric Pighin. 1998. Making faces. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '98)*. Association for Computing Machinery, New York, NY, USA, 55–66. https://doi.org/10.1145/280814.280822

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Ryan Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul E. Debevec, and Shahram Izadi. 2019. The relightables: volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.* 38, 6 (2019), 217:1–217:19. https://doi.org/10.1145/3355089.3356571

Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. 2024. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. *arXiv preprint arXiv:2402.10491* (2024).

Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. 2025. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In *European Conference on Computer Vision*. Springer, 39–55.

Moayed Haji-Ali, Guha Balakrishnan, and Vicente Ordonez. 2024. ElasticDiffusion: Training-free Arbitrary Size Image Generation through Global-Local Content Separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6603–6612.

Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. 2019. Recurrent back-projection network for video super-resolution. In *CVPR*. 3897–3906.

Jon Hasselgren, Jacob Munkberg, Marco Salvi, Anjul Patney, and Aaron Lefohn. 2020. Neural Temporal Adaptive Sampling and Denoising. *Computer Graphics Forum* (2020). https://doi.org/10.1111/cgf.13919

Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. 2024b. VEnhancer: Generative Space-Time Enhancement for Video Generation. *arXiv preprint arXiv:2407.07667* (2024).

Mingming He, Pascal Clausen, Ahmet Levent Tasel, Li Ma, Oliver Pilarski, Wenqi Xian, Laszlo Rikker, Xueming Yu, Ryan Burgert, Ning Yu, and Paul E. Debevec. 2024a. DifFRelight: Diffusion-Based Facial Performance Relighting. In *SIGGRAPH Asia 2024 Conference Papers, SA 2024, Tokyo, Japan, December 3-6, 2024*, Takeo Igarashi, Ariel Shamir, and Hao (Richard) Zhang (Eds.). ACM, 11:1–11:12. https://doi.org/10.1145/3680528.3687644

Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. 2024c. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*.

Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel J. Brostow. 2018. Deep blending for free-viewpoint image-based rendering. *ACM Trans. Graph.* 37, 6 (2018), 257. https://doi.org/10.1145/3272127.3275084

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. 2023. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*. PMLR, 13213–13232.

Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. 2024. FouriScale: A Frequency Perspective on Training-Free High-Resolution Image Synthesis. *arXiv preprint arXiv:2403.12963* (2024).

Yan Huang, Wei Wang, and Liang Wang. 2017. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE TPAMI* 40, 4 (2017), 1015–1028.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2023. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982* (2023).

Juno Hwang, Yong-Hyun Park, and Junghyo Jo. 2024. Upsample guidance: Scale up diffusion models without training. *arXiv preprint arXiv:2404.01709* (2024).

Mustafa Isik, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Trans. Graph.* 42, 4 (2023), 160:1–160:12. https://doi.org/10.1145/3592415

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. 2024. Pyramidal Flow Matching for Efficient Video Generative Modeling. (2024).

Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. 2023. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems* 36 (2023), 70847–70860.

T. Kanade, P. Rander, and P.J. Narayanan. 1997. Virtualized reality: constructing virtual worlds from real scenes. *IEEE MultiMedia* 4, 1 (1997), 34–47. https://doi.org/10.1109/93.580394

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 2024. 3D Gaussian Splatting as Markov Chain Monte Carlo. In *Advances in Neural Information Processing Systems (NeurIPS)*. Spotlight Presentation.

Georgios Kopanas, Julien Philip, Thomas Leimkühler, and George Drettakis. 2021. Point-Based Neural Rendering with Per-View Optimization. *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering)* 40, 4 (June 2021). http://www-sop.inria.fr/reves/Basilic/2021/KPLD21

Black Forest Labs. 2024. *FLUX.1 : An advanced state-of-the-art generative deep learning model.* Technical Report. Black Forest Labs. https://flux1.io/

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. 2023. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems* 36 (2023), 50648–50660.

Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C. Buehler, Otmar Hilliges, and Thabo Beeler. 2022. EyeNeRF: A Hybrid Representation for Photorealistic Synthesis, Animation and Relighting of Human Eyes. *ACM Trans. Graph.* 41, 4, Article 166 (jul 2022), 16 pages. https://doi.org/10.1145/3528223.3530130

Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. 2020. Mucan: Multi-correspondence aggregation network for video super-resolution. In *ECCV*. Springer, 335–351.

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. 2024. Spacetime Gaussian Feature Splatting for Real-Time Dynamic View Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8508–8520.

Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. 2024. Vrt: A video restoration transformer. *IEEE TIP* (2024).

Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. 2022. Recurrent video restoration transformer with guided deformable attention. *NeurIPS* 35 (2022), 378–393.

Haotong Lin, Sida Peng, Zhen Xu, Tao Xie, Xingyi He, Hujun Bao, and Xiaowei Zhou. 2023b. Im4D: High-Fidelity and Real-Time Novel View Synthesis for Dynamic Scenes. *CoRR* abs/2310.08585 (2023). https://doi.org/10.48550/ARXIV.2310.08585 arXiv:2310.08585

Mingbao Lin, Zhihang Lin, Wengyi Zhan, Liujuan Cao, and Rongrong Ji. 2024. CutDiffusion: A Simple, Fast, Cheap, and Strong Diffusion Extrapolation Method. *arXiv preprint arXiv:2404.15141* (2024).

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. 2023a. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070* (2023).

Songhua Liu, Weihao Yu, Zhenxiong Tan, and Xinchao Wang. 2024. LinFusion: 1 GPU, 1 Minute, 16K Image. (2024). arXiv:2409.02097 [cs.CV]

Stephen Lombardi, Tomas Simon, Jason M. Saragih, Gabriel Schwartz, Andreas M. Lehrmann, and Yaser Sheikh. 2019. Neural volumes: learning dynamic renderable volumes from images. *ACM Trans. Graph.* 38, 4 (2019), 65:1–65:14. https://doi.org/10.1145/3306346.3323020

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhöfer, Yaser Sheikh, and Jason M. Saragih. 2021. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.* 40, 4 (2021), 59:1–59:13. https://doi.org/10.1145/3450626.3459863

Jundan Luo, Duygu Ceylan, Jae Shin Yoon, Nanxuan Zhao, Julien Philip, Anna Frühstück, Wenbin Li, Christian Richardt, and Tuanfeng Y. Wang. 2024. IntrinsicDiffusion: Joint Intrinsic Layers from Latent Diffusion Models. In *SIGGRAPH 2024 Conference Papers*. https://doi.org/10.1145/3641519.3657472

Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V. Sander. 2021. Deblur-NeRF: Neural Radiance Fields from Blurry Images. *arXiv preprint arXiv:2111.14292* (2021).

Abhimitra Meka, Rohit Pandey, Christian Häne, Sergio Orts-Escolano, Peter Barnum, Philip Davidson, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe LeGendre, Wan-Chun Ma, Ryan S. Overbeck, Thabo Beeler, Paul E. Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Ryan Fanello. 2020. Deep relightable textures: volumetric performance capture with neural rendering. *ACM Trans. Graph.* 39, 6 (2020), 259:1–259:21. https://doi.org/10.1145/3414685.3417814

Marko Mihajlovic, Sergey Prokudin, Siyu Tang, Robert Maier, Federica Bogo, Tony Tung, and Edmond Boyer. 2024. SplatFields: Neural Gaussian Splats for Sparse 3D and 4D Reconstruction. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part II* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 313–332. https://doi.org/10.1007/978-3-031-72627-9_18

Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. 2022. NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images. *CVPR* (2022).

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters* 20, 3 (2013), 209–212. https://doi.org/10.1109/LSP.2012.2227726

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (July 2022), 15 pages. https://doi.org/10.1145/3528223.3530127

Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021a. Nerfies: Deformable Neural Radiance Fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 5845–5854. https://doi.org/10.1109/ICCV48922.2021.00581

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021b. HyperNeRF: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.* 40, 6 (2021), 238:1–238:12. https://doi.org/10.1145/3478513.3480487

Philipp; Mildenhall Ben; Barron Jonathan T.; Martin-Brualla Ricardo Park, Keunhong; Henzler. 2023. CamP: Camera Preconditioning for Neural Radiance Fields. *ACM Trans. Graph.* (2023).

Sida Peng, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2023. Representing Volumetric Videos as Dynamic MLP Maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 4252–4262. https://doi.org/10.1109/CVPR52729.2023.00414

Julien Philip and Valentin Deschaintre. 2023. Floaters No More: Radiance Field Gradient Scaling for Improved Near-Camera Training. In *Eurographics Symposium on Rendering*, Tobias Ritschel and Andrea Weidlich (Eds.). The Eurographics Association. https://doi.org/10.2312/sr.20231122

Haonan Qiu, Shiwei Zhang, Yujie Wei, Ruihang Chu, Hangjie Yuan, Xiang Wang, Yingya Zhang, and Ziwei Liu. 2024. FreeScale: Unleashing the Resolution of Diffusion Models via Tuning-Free Scale Fusion. *arXiv preprint arXiv:2412.09626* (2024).

Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. 2024. Ultrapixel: Advancing ultra-high-resolution image synthesis to new peaks. *arXiv preprint arXiv:2407.02158* (2024).

Barbara Roessle, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. 2023. GANeRF: Leveraging Discriminators to Optimize Neural Radiance Fields. *ACM Trans. Graph.* 42, 6, Article 207 (nov 2023), 14 pages. https://doi.org/10.1145/3618402

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*. 10684–10695.

Darius Rückert, Linus Franke, and Marc Stamminger. 2022. Adop: Approximate differentiable one-pixel point rendering. *ACM Transactions on Graphics (ToG)* 41, 4 (2022), 1–14.

Runway AI, Inc. 2024. *Gen-3 Alpha.* https://runwayml.com/research/introducing-gen-3-alpha

Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. 2018. Frame-recurrent video super-resolution. In *CVPR*. 6626–6634.

Richard Shaw, Michal Nazarczuk, Jifei Song, Arthur Moreau, Sibi Catley-Chandar, Helisa Dhamo, and Eduardo Pérez-Pellitero. 2024. SWinGS: Sliding Windows for Dynamic 3D Gaussian Splatting. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LV* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 37–54. https://doi.org/10.1007/978-3-031-73001-6_3

Yuan Shen, Duygu Ceylan, Paul Guerrero, Zexiang Xu, Niloy J. Mitra, Shenlong Wang, and Anna Frühstück. 2024. SuperGaussian: Repurposing Video Models for 3D Super Resolution. In *European Conference on Computer Vision (ECCV)*.

Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. 2022. Rethinking alignment in video super-resolution transformers. *NeurIPS* 35 (2022), 36081–36093.

Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. 2024. 3DGStream: On-the-Fly Training of 3D Gaussians for Efficient Streaming of Photo-Realistic Free-Viewpoint Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20675–20685.

Zachary Teed and Jia Deng. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II* (Glasgow, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 402–419. https://doi.org/10.1007/978-3-030-58536-5_24

Jiayan Teng, Wendi Zheng, Ming Ding, Wenyi Hong, Jianqiao Wangni, Zhuoyi Yang, and Jie Tang. 2023. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. *arXiv preprint arXiv:2309.03350* (2023).

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new Metric for Video Generation.. In *DGS@ICLR*. OpenReview.net. http://dblp.uni-trier.de/db/conf/iclr/dgs2019.html#UnterthinerSKMM19

Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popovic. 2008. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.* 27, 3 (2008), 97. https://doi.org/10.1145/1360612.1360696

Daniel Vlasic, Pieter Peers, Ilya Baran, Paul E. Debevec, Jovan Popovic, Szymon Rusinkiewicz, and Wojciech Matusik. 2009. Dynamic shape capture using multi-view photometric stereo. *ACM Trans. Graph.* 28, 5 (2009), 174. https://doi.org/10.1145/1618452.1618520

Chao Wang, Krzysztof Wolski, Bernhard Kerbl, Ana Serrano, Mojtaba Bemana, Hans-Peter Seidel, Karol Myszkowski, and Thomas Leimkühler. 2024b. Cinematic Gaussians: Real-Time HDR Radiance Fields with Depth. In *Computer Graphics Forum*, Vol. 43. Blackwell-Wiley, 1–13.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. 2024d. Exploiting diffusion prior for real-world image super-resolution. *IJCV* (2024), 1–21.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. 2023. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103* (2023).

Yuehao Wang, Chaoyi Wang, Bingchen Gong, and Tianfan Xue. 2024a. Bilateral Guided Radiance Field Processing. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–13.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. 2024c. SinSR: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25796–25805.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024b. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20310–20320.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. 2025. Difix3D+: Improving 3D Reconstructions with Single-Step Diffusion Models. arXiv:2503.01774 [cs.CV]

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. 2024a. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*. 25456–25467.

Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. 2025. STAR: Spatial-Temporal Augmentation with Text-to-Video Models for Real-World Video Super-Resolution. *arXiv preprint arXiv:2501.02976* (2025).

Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. 2021. Temporal modulation network for controllable space-time video super-resolution. In *CVPR*. 6388–6397.

Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. 2022. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5438–5448.

Zexiang Xu, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi. 2019. Deep view synthesis from sparse photometric images. *ACM Trans. Graph.* 38, 4 (2019), 76:1–76:13. https://doi.org/10.1145/3306346.3323007

Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 2024a. 4K4D: Real-Time 4D View Synthesis at 4K Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 20029–20040. https://doi.org/10.1109/CVPR52733.2024.01893

Zhen Xu, Yinghao Xu, Zhiyuan Yu, Sida Peng, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. 2024b. Representing Long Volumetric Video with Temporal Gaussian Hierarchy. *ACM Trans. Graph.* 43, 6 (2024), 171:1–171:18. https://doi.org/10.1145/3687919

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. 2023. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469* (2023).

Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. 2024a. Motion-Guided Latent Diffusion for Temporally Consistent Real-world Video Super-resolution. (2024).

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).

Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. 2024c. Real-time Photorealistic Dynamic Scene Representation and Rendering with 4D Gaussian Splatting. *International Conference on Learning Representations (ICLR)*.

Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. 2025. gsplat: An open-source library for Gaussian splatting. *Journal of Machine Learning Research* 26, 34 (2025), 1–17.

Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. 2019. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*. 3106–3115.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. 2024. Mip-Splatting: Alias-free 3D Gaussian Splatting. *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

Xin Yuan, Jinoo Baek, Keyang Xu, Omer Tov, and Hongliang Fei. 2024. Inflation with Diffusion: Efficient Temporal Adaptation for Text-to-Video Super-Resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 489–496.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. 2024. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *NeurIPS* 36 (2024).

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. 2024. RGB↔X: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 75, 11 pages. https://doi.org/10.1145/3641519.3657445

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*. 586–595.

Yuehan Zhang and Angela Yao. 2024. RealViformer: Investigating Attention for Real-World Video Super-Resolution. *ECCV* (2024).

Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. 2024. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8281–8291.

Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, and Jingyi Yu. 2022. Human Performance Modeling and Rendering via Neural Animated Mesh. *ACM Trans. Graph.* 41, 6 (2022), 235:1–235:17. https://doi.org/10.1145/3550454.3555451

Qingping Zheng, Yuanfan Guo, Jiankang Deng, Jianhua Han, Ying Li, Songcen Xu, and Hang Xu. 2024. Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7571–7578.

Kun Zhou, Wenbo Li, Yi Wang, Tao Hu, Nianjuan Jiang, Xiaoguang Han, and Jiangbo Lu. 2023. NeRFLix: High-Quality Neural View Synthesis by Learning a Degradation-Driven Inter-Viewpoint MiXer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12363–12374.

Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. 2024. Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution. In *CVPR*. 2535–2545.