

Unwarping the Lens: A Physics-Grounded Approach to Video Glasses Removal

Radim Spetlik^{1*}, David Futschik², Radek Danecek²,
Feitong Tan², Ziqian Bai², Rohit Pandey², and Yinda Zhang²

¹ Czech Technical University in Prague, FEE, Czech Republic

² Google, USA, * work done while at Google

Abstract. High-fidelity removal of eyeglasses from video is a major challenge in facial attribute editing, as the underlying facial geometry is often obscured by complex refractive distortions and view-dependent specular reflections. While large-scale generative priors have shown promise in eyeglasses removal via static image inpainting, they often lack the structural constraints necessary to maintain identity, expression, and pose, leading to visible “identity drift” in both static images and dynamic sequences. In this paper, we propose a novel transfer framework that addresses the stochastic nature of generative priors. Our pipeline first extracts high-fidelity synthetic face images from a commercial-grade generative model (Nano Banana, Gemini 3 Pro Image), regularizes them via a three-stage structural filtering process to preserve identity, expression, and pose, and finally applies physically-based simulation of lens optics during training to provide diverse, paired data. This process transfers Nano Banana’s photo-realistic, multi-view knowledge into a specialized restoration architecture, *JFSnet* (Joint Feature-Spatial network). JFSnet integrates DINOv2-based semantic features with a convolutional decoder for spatial reconstruction, leveraging translation equivariance constraints to improve temporal consistency and high-frequency detail preservation. Evaluations on the curated Flickr-Faces-HQ (FFHQ) subset (12,163 images) show that our approach achieves high fidelity and structural accuracy, while maintaining inference speed of 27.68 FPS. In perceptual studies on CelebV-Text video sequences, our results are consistently preferred over diffusion and GAN-based baselines for ocular consistency, temporal stability, and overall restoration quality.

Keywords: Image Editing · Video Editing · Glasses Removal

1 Introduction

The growth of high-resolution video content has increased the demand for sophisticated facial attribute editing. While significant progress has been made in global style transfer and large-scale object manipulation [2, 11, 25], precise local editing remains a difficult task. Eyeglasses, specifically, are not just blocking layers; they are complex optical elements that introduce depth-dependent

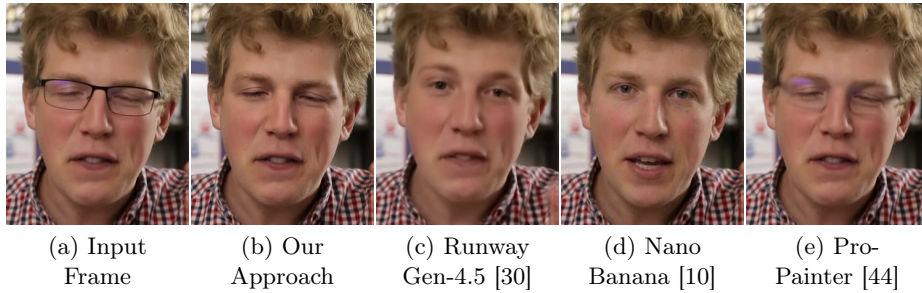


Fig. 1: Joint Feature-Spatial network removes glasses in videos producing results that are both temporally consistent and structurally faithful. Given an input video frame (a), our approach (b) effectively erases the glasses while preserving facial identity. In comparison, the commercial video editing model Runway Gen-4.5 [30] alters the subject’s expression and introduces blurriness (c). Similarly, the image-based editor Nano Banana [10] fails to preserve the expression when applied to video frames (d). The video inpainting method ProPainter [44] faces different limitations, resulting in incomplete glasses removal (e). In contrast, our method (b) ensures coherent removal and maintains high-fidelity structural integrity throughout the sequence. Source: CelebV-Text [37].

refractive distortions and view-dependent specular reflections. Standard generative inpainting models [29, 38] tend to struggle to separate these rare optical effects from the underlying facial geometry, which can lead to “identity drift” and structural inconsistencies in both static images and videos.

Existing approaches to glasses removal [19, 32] primarily rely on stochastic generative priors to hallucinate the occluded regions. However, in the presence of strong refractive lenses, the human visual system relies on stable geometric cues, such as the continuity of the head contour and the undistorted shape of the eyes, to maintain identity. When a model relies on black-box priors to reconstruct the eye area, it often produces results that look realistic in isolation but can fail to preserve the subject’s unique structural features, such as the exact expression, pose, and identity. Furthermore, applying these methods frame-by-frame in a naive manner leads to temporal flickering, as the generative process is unconstrained across the temporal domain.

In this work, we propose a novel transfer framework that addresses the stochastic nature of these generative priors. Our approach leverages large-scale models like Nano Banana (Gemini 3 Pro Image) [10], which provide photo-realistic results with multi-view consistency, but often struggle to preserve fine-grained identity, expression, and pose. To address these limitations, we propose improvements to both data curation and model architecture. On the side of data generation, we introduce a structured curation and augmentation pipeline: (i) we perform multi-view generative synthesis of eyeglass-free (“clean”) and glasses-wearing identities in sets of 13 poses and expressions using Nano Banana, (ii) we apply a structural filtering process — first identifying pose shifts via background L_1 error, then filtering gaze and identity drift via localized structural consistency

analysis, and finally utilizing a learned reconstruction threshold — to ensure a high-fidelity curated source, and (iii) we finally apply physically-based substitution of lens optics during training to provide diverse and realistic pairs. This process allows us to transfer the knowledge of large-scale generative models into a restoration architecture, *JFSnet*, to achieve inference speeds suitable for real-time video applications.

In contrast to latent diffusion approaches that are constrained by the manifold of a fixed pre-trained autoencoder, our method operates directly in pixel space. This allows JFSnet to preserve facial details that might be lost during latent reconstruction. By leveraging a self-supervised DINOv2-based encoder, we capture stable semantic priors that are invariant to local distortions, while a precision-focused convolutional decoder reconstructs fine facial features. To address the challenge of temporal instability without the high computational cost of multi-frame propagation or flow estimation, we enforce *translation equivariance* with a training constraint. By encouraging the network to commute with spatial shifts, we improve the coherence of restored features across frames, helping to suppress temporal flickering.

In summary, our contributions are as follows. We

- (i) formulate video glasses removal as the offline transfer of a generative prior into a deterministic, feed-forward model, and introduce a pipeline that combines curated generative face sets with physically-based augmentation of lens optics during training;
- (ii) present a framework utilizing *JFSnet*, a ViT-CNN architecture enabling identity preservation and temporal consistency;
- (iii) evaluate on the curated FFHQ subset, outperforming state-of-the-art generative and video inpainting baselines with inference speed of 27.68 FPS;
- (iv) demonstrate through perceptual studies that our approach is preferred across gaze preservation, temporal stability, removal quality, identity preservation, and sharpness.

Source code and dataset available at radimspetlik.github.io/unwarpingthelens/.

2 Related Work

2.1 Eyeglasses Removal

Early work in glasses removal primarily focused on static portrait editing using Generative Adversarial Networks (GANs). ByeglassesGAN [19] and subsequent work [23] leveraged GAN-based architectures to synthesize eyeglass-free ocular regions, with the latter utilizing 3D synthetic data to improve the restoration of cast shadows. More recently, video-specific approaches have emerged to address temporal consistency. V-LASIK [32] introduces identity-preserving constraints and color normalization to align restored regions with the rest of the face. Similarly, IP-FaceDiff [1] utilizes diffusion priors to handle the removal of eyewear in dynamic sequences. While these methods show promise, they often exhibit identity drift and local blurring because they rely on stochastic generative priors

that lack explicit geometric constraints. Our work differs by utilizing a physics-based simulation of lens optics during training, allowing the model to learn to invert the specific refractive distortions introduced by eyewear.

2.2 Generative Image and Video Editing

Advances in diffusion models have enabled broad image editing capabilities [5, 12, 35]. InstructPix2Pix [5] and LEDITS [35] allow for instruction-based and semantic editing, while video-to-video translation models [8, 15, 27] propagate edits across frames to maintain coherence. However, many of these methods operate in the latent space of pre-trained autoencoders, which introduce a bottleneck that inevitably sacrifices fine-grained spatial details like skin texture and eye geometry. Furthermore, while general-purpose models excel at global style transfer, they often struggle with precise local attribute removal, frequently failing to preserve the underlying facial structure.

Video inpainting and object removal methods [39, 42, 44] address occlusions by propagating information from adjacent frames. However, in the context of glasses removal, the “background” is the subject’s own eyes — critical features for identity that are often refracted or partially occluded rather than entirely missing. Generic inpainting approaches lack the specialized facial priors needed to accurately reconstruct these details, often leading to hallucinatory results that do not meet the bar of perceived geometric fidelity.

2.3 Identity Preservation and 3D-Aware Editing

Preserving identity is a core challenge in facial editing. Recent works utilize pre-trained Vision Transformers (ViTs) as semantic identity priors [26, 34], while 3D-aware methods leveraging Morphable Models (3DMMs) [3] or Neural Radiance Fields (NeRFs) [24] maintain consistency under varying viewpoints. Recent methods integrate physical optical transport — refraction, shadows, and specularities — into 3D morphable models and neural radiance fields to accurately model eyewear [20, 21, 40]. Similarly, 3D Gaussian Splatting enables high-fidelity dynamic avatars [22, 28, 31, 41], though these representations rarely support removal or refraction. While some explicitly regress 3D facial topology to handle occlusion [43], these representations often demand multi-view calibration or expensive optimization. Our approach bridges the gap between 2D translation and 3D geometric reasoning by incorporating physics-based refraction models, ensuring identity preservation without the need for per-scene 3D reconstruction or expensive overhead.

3 Method

We propose a novel transfer framework for consistent glasses removal that addresses the tendency of generative priors toward identity divergence and the introduction of structural inconsistencies (see Fig. 1). Our approach utilizes

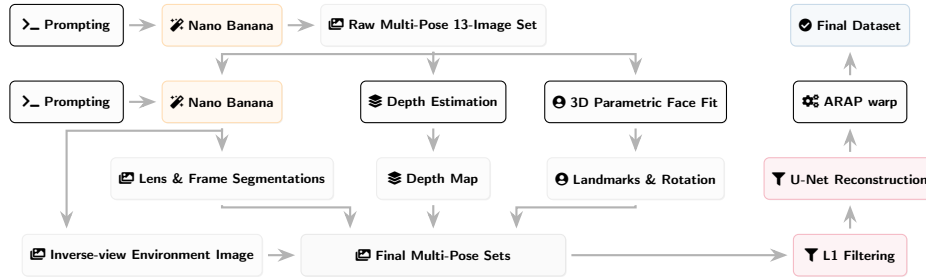


Fig. 2: Dataset creation pipeline. Raw multi-pose face sets are generated from the Nano Banana prior and enriched with segmentation masks, monocular depth, and parametric face model fits. These sets undergo multi-stage filtering (joint background/ocular *L1 Filtering*, followed by U-Net reconstruction) and an As-Rigid-As-Possible (ARAP) warp to align clean and glasses-wearing pairs, giving the final dataset.

a restoration network, *JFSnet*, which learns to invert complex optical effects while preserving identity, expression, and pose.

3.1 Generative Data Creation and Filtering

To support our training pipeline, we constructed a dataset of eyeglass-free (“clean”) and glasses-wearing face pairs using a generative workflow powered by Nano Banana (Gemini 3 Pro Image) [10] shown in Fig. 2. For each synthetic identity, we prompted the model to generate a set of 13 high-fidelity clean images capturing a diverse range of head poses and facial expressions. We further generated a standalone image of eyeglasses and a corresponding set of composed images where the subject wears the glasses in each pose, along with precise segmentation masks for the left lens, right lens, and the frames for every sample in the sequence. See prompts in Supplementary Sec. S1.2. This provided us with a comprehensive set of multi-pose and multi-expression pairs with ground-truth frame textures and pixel-accurate lens regions. A challenge in generative data synthesis is maintaining spatial alignment between the eyeglass-free and glasses-wearing pairs due to a number of effects, mainly the propensity of ViTs to overcompress spatial information and stochasticity in latent generative models. To address this, we used the composed Nano Banana images and their associated masks to define the precise geometric boundaries of the frames and lenses in each pose. Furthermore, we employ an *As-Rigid-As-Possible* (ARAP) [13] deformation mesh optimized with landmark constraints and a multi-scale photometric loss to align the clean face to the exact pose and facial configuration found in the glasses-wearing image (see Supplementary Sec. S1.9).

To ensure high data quality, we performed a three-stage hard-sampling-based filtering process. In the first two stages, we perform background and ocular L_1 filtering (*L1 Filtering* in Fig. 2) to filter out identities with significant shifts in head pose, gaze, or eye identity. This process effectively eliminates samples where the generative model introduced unwanted expression drift; such expression drift



Fig. 3: Refraction simulation. Physics-based lens distortion for -6 , -4 , -2 , and $+1$ diopters on a sample from our dataset. Our refraction model captures realistic magnification and minification while preserving facial identity and structure.

is exemplified by Nano Banana in Fig. 1 (d). This stage remains crucial because we down-weight the physics-based lens substitution for extreme head poses — specifically rotations approaching a profile view (beyond roughly three-quarters of the way from frontal to full profile) — where monocular depth-based simulation becomes less reliable and we rely on the generative model’s native outputs. In the third stage, we trained a lightweight U-Net on the remaining pairs to perform a preliminary restoration task, which serves as a quality proxy; identities with high reconstruction error were discarded as they typically exhibited residual misalignments or generative artifacts. This elaborate filtering resulted in a final training dataset of 1,860 identities exhibiting high structural and photometric consistency across the 13-image sets (see the Supplementary Sec. S1.1).

3.2 Physically-Grounded Training Augmentation

Although this dataset is rich in identities, the optical effects represented in generated data are far more constrained, likely following the generative model’s simplicity bias. To address the lack of large-scale paired data, we introduce a physics-based augmentation pipeline that simulates the optical properties of lenses on the curated generative face sets (see Fig. 3). This pipeline operates during training to provide a source of consistent and diverse training pairs.

Refraction Simulation (Fig. 3). To ensure realistic placement and distortion, we first identify the lens boundaries using precise segmentation masks provided by Nano Banana. The relative 3D pose of the head is estimated by fitting a 3D face parametric model [7] to the input frame to extract the global rotation $\mathbf{R}$ and translation $\mathbf{t}$ relative to the camera. We define the optical center of each lens by back-projecting the detected pupil landmarks using the depth map from a single-frame depth estimation model [36]. To resolve the scale ambiguity inherent in monocular depth estimation, we assume a standard adult head width to anchor the metric scale. The lenses are then positioned by projecting these metric 3D pupil coordinates 15mm along the head pose direction (the normal vector derived from $\mathbf{R}$). We assume a standard perspective camera model with a typical portrait focal length to initialize the 3D projection. The corrective lens is represented as

a physical 3D object defined by two spherical surfaces with radii R_1 and R_2 , positioned relative to the eye centers. These radii are derived from the target diopter strength D using Vogel’s Rule [4]: $D_1 = D/2 + 6.0$ and $D_2 = D - D_1$, where $D = (n - 1)(1/R_1 - 1/R_2)$. For each pixel within the lens boundary, a ray is cast and intersected with these surfaces. At each interface, the ray direction is updated via Snell’s law ($n \approx 1.5$). The resulting intersection with the facial surface defines the source coordinates for the dense warp field $\mathcal{W}$, simulating the depth-dependent magnification of prescription eyewear. Our rendering pipeline operates on high-resolution grids to preserve detail.

Reflection Simulation. Specular reflections $\mathcal{R}$ are simulated by leveraging Nano Banana to generate an inverse-view environment map, conditioned on the subject’s portrait, that approximates the scene behind the camera, sampled along the reflected view ray $\mathbf{r} = \mathbf{i} - 2(\mathbf{n}^\top \mathbf{i})\mathbf{n}$ ($\mathbf{i}$: incident view direction, $\mathbf{n}$: lens-surface normal). To achieve realistic highlights, we simulate reflections in High Dynamic Range (HDR), allowing for strong specularities of glass surfaces. We further augment these reflections by simulating a variable reflection tint and opacity, accounting for the diverse range of lens coatings and lighting conditions.

Training Input Formation. To provide a physically grounded yet photo-realistic training source, we define the observed image $I \in \mathbb{R}^{H \times W \times 3}$ as a composition of generative templates and optical simulations. Let G_{clean} and $G_{glasses}$ be the multi-pose, multi-expression sets from our curated dataset. During training, we synthesize the final input I by selecting a pose-expression pair and substituting the lens regions of $G_{glasses}$ with a physically simulated rendering of the corresponding clean face G_{clean} :

$$I = (1 - M_{lens}) \cdot G_{glasses} + M_{lens} \cdot (\mathcal{W}(G_{clean}) + \mathcal{R}) + \epsilon \quad (1)$$

where M_{lens} is a binary mask for the lens regions (provided by the left and right lens masks from the generative model), $\mathcal{W}(\cdot)$ represents the refractive warping operator, and $\mathcal{R}$ is the additive HDR specular reflection layer. This formulation ensures that the network learns to invert the refractive transformation $\mathcal{W}$ and handle additive reflections $\mathcal{R}$ while benefiting from the realistic frame textures and global illumination present in the generative model $G_{glasses}$. The network’s goal is to map this input I back to the original clean identity G_{clean} .

3.3 JFSnet Architecture

We utilize a restoration network (f_θ), referred to as *JFSnet*, designed to map the input image to the clean RGB face. As detailed in Fig. S2, *JFSnet* is a combination of standard architectural blocks: a pre-trained DINOv2 (ViT-L/14) encoder [26] and a ResNet-based convolutional decoder. To preserve representational power while adapting to the restoration task, we selectively fine-tune the deeper layers of the ViT backbone and fuse multi-level features into the decoder. Crucially, by incorporating a direct skip connection from the input image, we bypass the information bottleneck typical of latent models. This design allows

JFSnet to reconstruct high-frequency details that might fall outside standard generative manifolds, ensuring that the restored facial features remain faithful to the original input. As discussed in Sec. 4.9, this configuration was found to significantly outperform alternative architectures.

3.4 Training Objectives

The network f_θ is trained using a weighted combination of reconstruction and perceptual losses. Let $f_\theta(I)$ be the predicted clean face and G_{clean} be the ground truth clean face. The total loss $\mathcal{L}$ is defined as:

$$\begin{aligned} \mathcal{L} = & \lambda_{pix}\mathcal{L}_{pix} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{eye-pix}\mathcal{L}_{eye-pix} \\ & + \lambda_{eye-perc}\mathcal{L}_{eye-perc} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{temp}\mathcal{L}_{temp} \end{aligned} \quad (2)$$

Global Reconstruction and Perceptual. $\mathcal{L}_{pix}$ is the global L_1 distance between the predicted output and the ground truth. $\mathcal{L}_{perc}$ ensures high-level texture consistency across the full image, defined by the distance between Gram matrices of VGG-19 features.

Eye-specific Objectives. To restore features behind lenses, we utilize localized objectives centered on the center of gravity (CoG) of each lens mask within 64×64 pixel blocks. $\mathcal{L}_{eye-pix}$ is the localized L_1 reconstruction loss, while $\mathcal{L}_{eye-perc}$ is the corresponding Gram-matrix based perceptual loss.

Adversarial Loss. $\mathcal{L}_{adv}$ incorporates feedback from a PatchGAN-based critic [14] to ensure realistic texture generation, encouraging the network to produce sharp, high-frequency details. The critic is trained in the standard GAN min-max setting [9], though we omit this formulation for brevity.

Temporal Consistency. To ensure temporally stable results on video data, we enforce translation equivariance during training. For each training sample, we simulate motion by applying a random spatial shift Δs to the input image I . The temporal consistency loss $\mathcal{L}_{temp}$ is defined as:

$$\mathcal{L}_{temp} = \|f_\theta(\mathcal{T}_{\Delta s}(I)) - \mathcal{T}_{\Delta s}(f_\theta(I))\|_1 \quad (3)$$

where $\mathcal{T}_{\Delta s}(\cdot)$ represents a translation by Δs . By forcing f_θ to commute with small spatial translations, we ensure that high-frequency artifacts and identity-defining features move coherently with the subject’s motion. This constraint encourages the model to learn a mapping that is robust to spatial misalignments and resists flickering across frames.

4 Experiments

In order to assess the efficacy of our proposed approach, we evaluate the task of glasses removal from images and consistent removal from videos. Our experiments are designed to validate the effectiveness of the physics-based synthesis pipeline and the *JFSnet* architecture in handling optical distortions and maintaining temporal stability.

4.1 Experimental Setup

Implementation Details. We implement our framework using JAX/Flax, training the models on high-resolution images resized to 518×518 pixels. We use the AdamW optimizer with a base learning rate of 3×10^{-4} for the *JFSnet* and *Adversarial Critic*. The pre-trained DINOv2 encoder is fine-tuned with a lower learning rate of 1×10^{-5} for its last four layers, while the rest of the backbone is kept frozen. Training is performed on TPU/GPU accelerators for approximately 400,000 batch iterations. For the physics-based augmentation, we sample diopter powers ranging from -6.0 to $+1.0$ with an assumed refractive index of $n \approx 1.5$. Since negative diopters (minification) introduce visible background artifacts into the eye area while positive diopters (magnification) are harder to distinguish, we focus on minification in our qualitative analysis (see Fig. 4) and simulation. Specular reflections are simulated via sphere mapping with high-dynamic-range (HDR) emulation and randomized anti-reflective (AR) coating effects. To mitigate aliasing, we compute a Level of Detail (LOD) map based on the Jacobian of the ray divergence, enabling interpolation between sharp and pre-blurred background images.

Data Augmentation Pipeline. We use our curated dataset of 1,860 synthetic identities (24,180 image pairs) with on-the-fly augmentation applied during training. By randomizing specular reflection intensity and tint to simulate diverse lens coatings, the model encounters a wide variety of optical distortions while remaining anchored to the eyeglass-free ground truth.

4.2 Dataset Generation

We constructed our training dataset using the generative workflow and filtering process detailed in Sec. 3.1. This process ensures high structural and photometric consistency across the multi-view generative face sets, providing a robust foundation for the subsequent physics-based augmentation. See Supplementary Sec. S1.1 for a qualitative sample of a single face set.

4.3 Datasets

We utilize the following datasets exclusively for qualitative and quantitative evaluation; neither dataset was used during the training phase of our models. **Flickr-Faces-HQ (FFHQ)** [16] consists of 70,000 high-resolution portrait images with significant variation in demographic factors and backgrounds. For our evaluation, we manually curated images containing eyewear from this dataset, classifying them into two distinct categories based on ocular visibility: **(i)** individuals wearing corrective glasses with visible eye regions, and **(ii)** individuals wearing sunglasses or eyewear with significant darkening and specular reflections. This process yielded a refined evaluation set containing 12,163 images with clear glasses and 2,176 images in the sunglasses category. The remaining images in the dataset ($\approx 55,000$) are utilized as the “without eyewear” reference

category. **CelebV-Text** [37] is a large-scale video dataset capturing dynamic facial expressions. For our qualitative evaluation and perceptual study (Sec. 4.7), we manually curated a subset of 60 video clips featuring subjects with clearly visible eyeglasses.

4.4 Baselines

We evaluate our method against representative approaches across three categories. For general video editing, we compare against TokenFlow [8], RAVE [15], and IP-FaceDiff [1]. In the domain of video inpainting, we include ProPainter [44], Flow-Guided Transformer (FGT) [42], Inpaint Anything [38], and STTN [39]. We also evaluate image-based facial editing methods including Take-Off-Eyeglasses (TOE) [23], LEDITS [35], and InstructPix2Pix [5].

Additionally, Runway Gen-4.5 is a private, commercial video generation model with significant inference costs, leading us to restrict its evaluation to the human preference study. For diffusion-based methods such as TokenFlow, RAVE, LEDITS, and IP-FaceDiff, achieving high-fidelity results is often sensitive to parameter tuning and the stability of the underlying inversion process (e.g., DDIM inversion [33]). As finding viable configurations exceeded the practical scope of our evaluation, we focused our human preference study on the examples where baseline methods achieved competitive scores on the FFHQ dataset.

4.5 Flickr-Faces-HQ Qualitative Results

We present a qualitative comparison on the FFHQ dataset in Fig. S4. The generative models, such as Nano Banana [10], often reconstruct the eye region with plausible but inaccurate details, leading to subtle identity shifts or gaze changes.

LEDITS [35] struggles to fully remove frame artifacts, whereas our method preserves facial textures and gaze and compensates for the spatial minification of strong refractive lenses (see Fig. 3). See the Supplementary Sec. S1.11 for more results and Sec. S1.12 for failure cases.

4.6 CelebV-Text Qualitative Results

The qualitative results for video sequences are shown in Fig. 5. Large-scale video-to-video model Runway Gen-4.5 [30] removes reflections, but often fails to preserve the subject’s gaze and eye identity. Video inpainting methods ProPainter [44] and FGT [42] are primarily designed for object removal and lack the facial priors needed to handle the complex shadows and reflections cast by eyeglasses. These methods often leave behind ghosting artifacts. By contrast, our method leverages its physics-based training, providing temporally stable results that maintain the geometric integrity of the face. Full videos are available in the supplement.

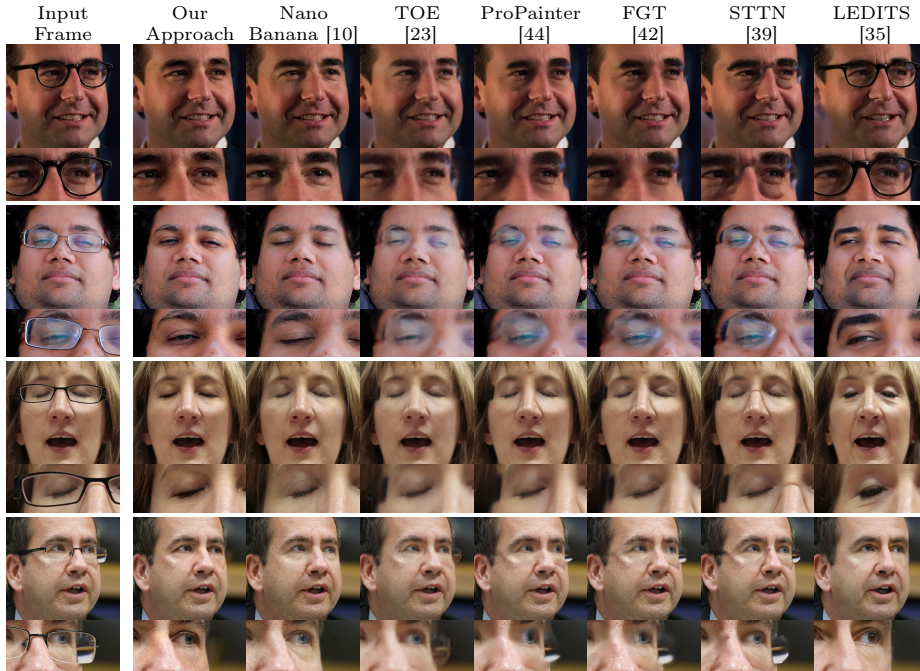


Fig. 4: Qualitative comparison on high-refraction samples from FFHQ [17]. Close-up eye regions are shown for each subject. While Nano Banana [10] removes reflections, it alters structural details (e.g., gaze shift in row 1, eye closure in row 2). Diffusion-based LEDITS [35] exhibits identity drift or fails to remove glasses (row 1). Video inpainting methods [39, 42, 44] often fail to recover underlying facial structures. Our method effectively removes reflections and corrects refraction while preserving fine facial details and maintaining high structural integrity. See Sec. 4.5 for more details.

4.7 Perceptual Study

To further evaluate our method, we conducted a perceptual study consisting of two surveys focused on video and image glasses removal. For the video survey, we utilized 60 clips from the CelebV-Text dataset, while the image survey used 60 high-resolution portraits from the FFHQ dataset. We invited 37 participants to compare our results against 4 other representative methods, resulting in a total of 5 methods per question. In each survey, participants were asked to rank the methods based on specific criteria derived from the qualitative characteristics of the task. For the video survey, the evaluation axes included gaze consistency, temporal stability, and overall restoration quality. The image survey focused on gaze preservation, removal completeness, identity preservation, sharpness, and overall removal quality. These criteria allow for a granular assessment of how well each method handles both the physical restoration of ocular details and the generative consistency of the face. The collected preferences are visualized in the heatmap of Fig. 6. Our method is consistently preferred across both temporal and

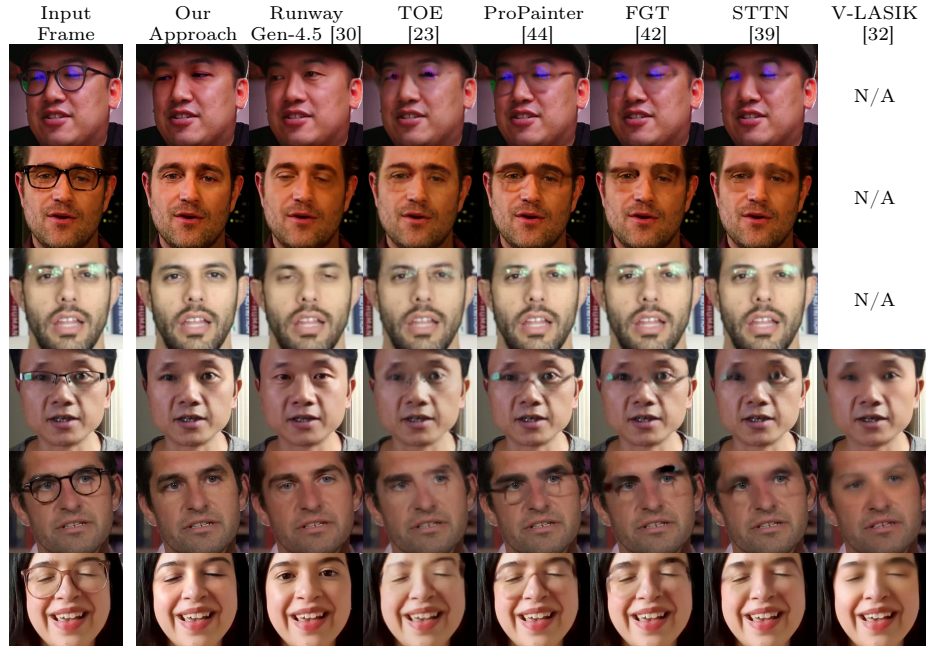


Fig. 5: Qualitative comparison on CelebV-Text [37]. Runway Gen-4.5 removes reflections but changes the gaze and eye identity. V-LASIK causes blur and changes eye shape. TOE struggles with strong reflections. ProPainter, FGT, and STTN only perform inpainting, leaving behind shadows, reflections, or parts of the glasses. Our Joint Feature-Spatial network preserves gaze and fine facial details. See Sec. 4.6.

spatial dimensions. While Nano Banana achieves high scores for completeness due to its powerful generative prior, it is consistently outperformed in gaze and identity preservation where its unconstrained outputs diverge from the subject’s original features. Temporal metrics are in the Supplementary Sec. S1.4.

4.8 Quantitative Evaluation

We assess the performance of the methods across several dimensions — gaze preservation (head pose, pupil position, and ocular structure), visual realism, structural fidelity, and computational efficiency.

Gaze preservation is quantified via the **Pupil Displacement error** (in px), which measures the spatial shift of iris centers between the input and restored output. To avoid evaluation bias, we detect pupils using the *MediaPipe FaceMesh* model [18], which is distinct from the models used in our data curation pipeline. We calculate the Euclidean distance in normalized coordinates and scale the average displacement by 512 for readability. Note that while physical refraction correction necessitates a minor spatial shift of the ocular features, a low displacement indicates that the model avoids the significant gaze drift common in

	Video Survey			Image Survey				
	Our method is preferred in			Our method is preferred in				
Nano Banana [10]	97%	100%	100%	81%	42%	75%	55%	74%
LEDITS [35]	100%	100%	96%	90%	85%	98%	90%	96%
FGT [42]	83%	96%	100%	80%	92%	82%	98%	97%
TOE [23]	91%	97%	98%	85%	96%	89%	100%	99%
V-LASIK [32]	88%	88%	100%	N/A	N/A	N/A	N/A	N/A
STTN [39]	83%	87%	100%	81%	100%	77%	92%	98%
ProPainter [44]	76%	76%	96%	84%	93%	84%	96%	98%
Runway Gen-4.5 [30]	74%	59%	75%	N/A	N/A	N/A	N/A	N/A
	when participants consider			when participants consider				
	Gaze consis- tency	Temporal consis- tency	Overall quality	Gaze preserva- tion	Removal complete- ness	Identity preserva- tion	Sharpness	Overall quality

Fig. 6: Preference study on CelebV-Text [37] and FFHQ [17]. Our approach is consistently preferred across both video and image surveys. The heatmap shows the percentage of participants who preferred our result over the baseline for each evaluation criterion. See Sec. 4.7 for more details.

generative baselines. Inpaint Anything [38] achieves a competitive pupil error as it ideally avoids ocular modification, but is prone to generating artifacts when segmentation fails. Conversely, Nano Banana [10] suffers from significant gaze drift despite serving as our primary data source.

To evaluate the visual realism and quality of the generated results, we employ the *Fréchet Inception Distance* (**FID**), which measures the statistical distance between the feature distributions of real and synthesized images to capture distribution-level fidelity.

Structural and geometric stability is measured by the **Landmark L2 Error** (L2), the Euclidean distance between facial landmarks in the input images (with glasses) and the corresponding landmarks in the restored outputs. This metric ensures that the restoration process does not introduce unwanted spatial distortions or identity-altering pose shifts. We empirically found that ArcFace [6] does not perform well for identity preservation analysis in our case (see Supp. Sec. S1.6). Finally, we report the *Inference Speed* in frames per second (FPS) to evaluate the practical utility of each method for video applications. We use the results of our perceptual study (Sec. 4.7) to measure temporal stability.

Results on FFHQ. We evaluate our method on the curated FFHQ evaluation set, primarily focusing on the clear glasses category where ocular details remain partially visible. As shown in Table 1, our approach achieves an FID of 0.379 and a landmark L2 error of 0.632. For the FID evaluation, we take the real images from the curated FFHQ “clear glasses” subset (12,163 images), resize

[†] Authors of V-LASIK [32] do not provide code and did not respond to our queries for evaluation; our comparison with their method is limited to the videos available on their project website.

Table 1: Quantitative evaluation on the Flickr-Faces-HQ (FFHQ) dataset [16]. We compare our proposed approach against several state-of-the-art image-based, video editing, and video inpainting methods using the Fréchet Inception Distance (FID) ↓, Pupil Displacement ↓, landmark L2 error (L2) ↓, and inference speed (FPS) ↑. Best results in **bold**. †

Method	FID ↓	Pupil (px) ↓	L2 (px) ↓	FPS ↑
Nano Banana [10]	0.389	2.854 ± 2.387	0.699 ± 0.462	1.29
Inpaint Anything [38]	0.391	2.259 ± 2.295	0.664 ± 0.902	1.84
InstructPix2Pix [5]	1.089	5.541 ± 6.522	8.038 ± 18.293	0.05
LEDITS [35]	1.953	3.905 ± 2.999	1.527 ± 2.437	0.08
Take-Off-Eyeglasses [23]	3.256	2.593 ± 1.696	0.877 ± 0.773	13.31
IP-FaceDiff [1]	3.832	17.607 ± 8.718	21.257 ± 27.171	0.04
RAVE [15]	11.559	3.577 ± 2.373	2.356 ± 2.773	0.16
TokenFlow [8]	2.430	3.689 ± 2.812	1.905 ± 1.678	0.04
Flow-Guided Transformer [42]	0.569	3.078 ± 2.714	0.965 ± 0.952	1.28
ProPainter [44]	0.408	2.684 ± 2.838	0.947 ± 1.095	8.77
Ours	0.379	2.249 ± 1.957	0.632 ± 0.271	27.68

them to 518×518 , run our removal pipeline, and compare the resulting distribution of restored images against a size-matched subset of the FFHQ “without eyewear” category. This distribution-level comparison validates that our method correctly restores facial regions to a state statistically indistinguishable from natural portraits without eyewear. The sunglasses category is in the Supplementary Sec. S1.3. Paired metrics and retrained baselines in the Supplementary Sec. S1.5.

4.9 Architectural Ablation

We compare the JFSnet configuration against standard baseline architectures to validate our choice of combining semantic and spatial blocks. Our evaluation includes a pure Vision Transformer (ViT-L/14), a standard U-Net with a ResNet-34 backbone, and the proposed JFSnet (see the Supplementary Sec. S1.7).

4.10 Ablation Study

To validate our design choices, we investigate the impact of the core components of our pipeline, as summarized in the Supplementary Table S8.

Our results demonstrate that both reflection and refraction simulations are critical for achieving low FID scores; removing either component leads to a degradation. While the temporal consistency loss (λ_{temp}) yields the same FID score for static images, it is included for its role in temporal stability; the combination of λ_{temp} , rigorous dataset filtering, and ARAP warping results in an FID of 0.379. The dataset filtering process—which includes L_1 thresholding and hard-sample mining—ensures the quality of the generative pairs, while the ARAP warping

(see Supplementary Sec. S1.9) accounts for residual pose shifts to maintain pixel-perfect alignment. Unfreezing the last four layers of the DINOv2 encoder provides a measurable boost in performance. Finally, the localized eye-specific losses ($\lambda_{\text{eye-pix}}$ and $\lambda_{\text{eye-perc}}$) enforce recovering details in the eye region.

5 Conclusion

In this paper, we presented a physics-grounded generative-to-deterministic transfer framework for consistent eyeglasses removal from video, bridging the gap between large-scale generative priors and physical restoration. By training our *JFSnet* architecture on a generative-to-physical synthesis pipeline, we achieved state-of-the-art visual fidelity (FID 0.379) while maintaining real-time performance at 27.68 FPS. Our approach leverages multi-level feature fusion and enforced translation equivariance to preserve gaze and identity across dynamic sequences, significantly outperforming existing baselines in perceptual studies.

Acknowledgments

The research reported in this paper has been partly funded by BMK, BMAW, and the State of Upper Austria in the frame of the SCCH competence center INTEGRATE [Project 1.6 TFI (Transferable Intelligence)] part of the FFG COMET Competence Centers for Excellent Technologies Programme, and by the Czech Technical University in Prague grant No. SGS26/074/OHK3/1T/13.

References

1. Anand, T., Garg, A., Mitra, K.: IP-FaceDiff: Identity-Preserving Facial Video Editing with Diffusion. pp. 248–258 (2025)
2. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Li, Y., Michaeli, T., others: Lumiere: A Space-Time Diffusion Model for Video Generation. In: SIGGRAPH Asia 2024 Conference Papers (2024)
3. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194. ACM Press/Addison-Wesley Publishing Co. (1999). <https://doi.org/10.1145/311535.311556>
4. Brooks, C.W., Borish, I.M.: System for Ophthalmic Dispensing. Butterworth-Heinemann (2007). <https://doi.org/10.1016/B978-0-7506-7480-5.X5001-1>
5. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to follow image editing instructions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
7. Egger, B., Smith, W.A.P., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., Theobalt, C., Blanz, V., Vetter, T.: 3D Morphable Face Models—Past, Present, and Future. ACM Trans. Graph. **39**(5), 157:1–157:38 (Jun 2020). <https://doi.org/10.1145/3395208>

8. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: TokenFlow: Consistent Diffusion Features for Consistent Video Editing (2023)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. pp. 2672–2680 (2014)
10. Google: Gemini 3 Pro Image - Model Card. Model Card, Google (2025), <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Image-Model-Card.pdf>, accessed: 2026-06-29
11. Google: Veo2 (2025), <https://veo2.ai/>
12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-Prompt Image Editing with Cross-Attention Control. In: *International Conference on Learning Representations* (2023)
13. Igarashi, T., Moscovich, T., Hughes, J.F.: As-rigid-as-possible shape manipulation. *ACM Trans. Graph.* **24**(3), 1134–1141 (2005). <https://doi.org/10.1145/1073204.1073323>
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1125–1134 (2017)
15. Kara, O., Kurtkaya, B., Yesiltepe, H., Rehg, J.M., Yanardag, P.: RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diffusion Models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024)
16. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4401–4410 (2019)
17. Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**(12), 4217–4228 (2021). <https://doi.org/10.1109/TPAMI.2020.2970919>
18. Kartynnik, Y., Ablavatski, A., Grishchenko, I., Grundmann, M.: Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs. In: *CVPR Workshop on Computer Vision for Augmented and Virtual Reality 2019* (2019)
19. Lee, Y.H., Lai, S.H.: ByeGlassesGAN: Identity Preserving Eyeglasses Removal for Face Images (Aug 2020). <https://doi.org/10.48550/arXiv.2008.11042>, <http://arxiv.org/abs/2008.11042>, arXiv:2008.11042 [cs]
20. Li, G., Meka, A., Mueller, F., Buehler, M.C., Hilliges, O., Beeler, T.: EyeNeRF: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Trans. Graph.* **41**(4), 166:1–166:16 (2022). <https://doi.org/10.1145/3528223.3530130>
21. Li, J., Saito, S., Simon, T., Lombardi, S., Li, H., Saragih, J.: MEGANE: Morphable Eyeglass and Avatar Network. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 12769–12779 (2023). <https://doi.org/10.1109/CVPR52729.2023.01228>
22. Liang, H., Ge, Z., Majee, S., Tiwari, A., Godaliyadda, G.M.D., Veeraraghavan, A., Balakrishnan, G.: FastAvatar: Instant 3D Gaussian Splatting for Faces from Single Unconstrained Poses (2025). <https://doi.org/10.48550/arXiv.2508.18389>
23. Lyu, J., Wang, Z., Xu, F.: Portrait eyeglasses and shadow removal by leveraging 3d synthetic data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3429–3439 (2022)

24. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**(1), 99–106 (2021). <https://doi.org/10.1145/3503250>
25. OpenAI: SORA (2025), <https://sora.com>
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2023)
27. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: FateZero: Fusing attentions for zero-shot text-based video editing. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 15932–15942 (2023)
28. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 20299–20309 (2024). <https://doi.org/10.1109/CVPR52733.2024.01919>
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. pp. 10684–10695 (2022)
30. Runway AI, Inc.: Runway Gen-4.5: High-Fidelity Text-to-Video and In-Context Editing (2025), <https://runwayml.com/research/introducing-runway-gen-4.5>, accessed: 2026-06-29
31. Serifi, G., Buehler, M.C.: HyperGaussians: High-Dimensional Gaussian Splatting for High-Fidelity Animatable Face Avatars. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2026)
32. Shalev-Arkushin, R., Azulay, A., Halperin, T., Richardson, E., Bermano, A.H., Fried, O.: V-LASIK: Consistent Glasses-Removal from Videos Using Synthetic Data. In: *ICLR 2025 Workshop on Will Synthetic Data Finally Solve the Data Access Problem?* (2025). <https://doi.org/10.48550/arXiv.2406.14510>
33. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models (2021)
34. Spetlik, R., Futschik, D., Sýkora, D.: StructuReiser: A Structure-preserving Video Stylization Method. *Computer Graphics Forum* **44**(4), e70161 (2025). <https://doi.org/10.1111/cgf.70161>
35. Tsaban, L., Passos, A.: LEDITS: Real image editing with DDPM inversion and semantic guidance. *arXiv preprint arXiv:2307.00522* (2023)
36. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. pp. 10371–10381 (2024)
37. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: CelebV-Text: A Large-Scale Facial Text-Video Dataset. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14805–14814 (2023)
38. Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., Chen, Z.: Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790* (2023)
39. Zeng, Y., Fu, J., Chao, H.: Learning Joint Spatial-Temporal Transformations for Video Inpainting. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) *Computer Vision – ECCV 2020*. pp. 528–543. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58517-4_31
40. Zhan, Y., Nobuhara, S., Nishino, K., Zheng, Y.: NeRFrac: Neural Radiance Fields through Refractive Surface. In: *2023 IEEE/CVF International Conference on Computer Vision*. pp. 18356–18366 (2023). <https://doi.org/10.1109/ICCV51070.2023.01687>

41. Zhang, D., Liu, Y., Lin, L., Zhu, Y., Chen, K., Qin, M., Li, Y., Wang, H.: HRA-vatar: High-Quality and Relightable Gaussian Head Avatar. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 26285–26296 (2025)
42. Zhang, K., Fu, J., Liu, D.: Flow-guided transformer for video inpainting. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 74–90 (2022)
43. Zhao, D., Qi, Y.: Generative Landmarks Guided Eyeglasses Removal 3D Face Reconstruction. In: MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part II. pp. 109–120. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-030-98355-0_10
44. Zhou, S., Li, C., Chan, K.C., Loy, C.C.: ProPainter: Improving Propagation and Transformer for Video Inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10477–10486 (2023)