DipGuava: Disentangling Personalized Gaussian Features for 3D Head Avatars from Monocular Video

Jeonghaeng Lee[1], **Seok Keun Choi**[1], **Zhixuan Li**[2], **Weisi Lin**[2], **Sanghoon Lee**[1]
[1]Yonsei University, Korea
[2]Nanyang Technological University, Singapore
leedoright@yonsei.ac.kr, csg@yonsei.ac.kr, zhixuan.li@ntu.edu.sg, wslin@ntu.edu.sg, slee@yonsei.ac.kr

Abstract

While recent 3D head avatar creation methods attempt to animate facial dynamics, they often fail to capture personalized details, limiting realism and expressiveness. To fill this gap, we present **DipGuava** (Disentangled and Personalized Gaussian UV Avatar), a novel 3D Gaussian head avatar creation method that successfully generates avatars with personalized attributes from monocular video. DipGuava is the first method to explicitly disentangle facial appearance into two complementary components, trained in a structured two-stage pipeline that significantly reduces learning ambiguity and enhances reconstruction fidelity. In the first stage, we learn a stable geometry-driven base appearance that captures global facial structure and coarse expression-dependent variations. In the second stage, the personalized residual details not captured in the first stage are predicted, including high-frequency components and nonlinearly varying features such as wrinkles and subtle skin deformations. These components are fused via dynamic appearance fusion that integrates residual details after deformation, ensuring spatial and semantic alignment. This disentangled design enables DipGuava to generate photorealistic, identity-preserving avatars, consistently outperforming prior methods in both visual quality and quantitative performance, as demonstrated in extensive experiments.

1. Introduction

Photorealistic 3D head avatars enable diverse applications such as VR, gaming, and telepresence, demanding efficient methods to create high-quality personalized avatars from monocular videos. Recent advancements, particularly Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) and 3D Gaussian splatting (3DGS) (Kerbl et al. 2023), have enabled highly photorealistic and animatable head avatar creation. A common strategy is to use 3D Morphable Models (3DMMs) (Paysan et al. 2009; Wang et al. 2022; Gerig et al. 2018; Li et al. 2017) as a stable geometric prior and as a shared latent space for expression and pose, making them suitable for animating diverse subjects. However, despite their progress, these approaches still struggle to capture the high-frequency, personalized attributes that define an individual's unique appearance, often due to their reliance on coarse geometric priors or entangled representations. As a result, they fall short in reproducing the subject-specific details essential for truly personalized avatars.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

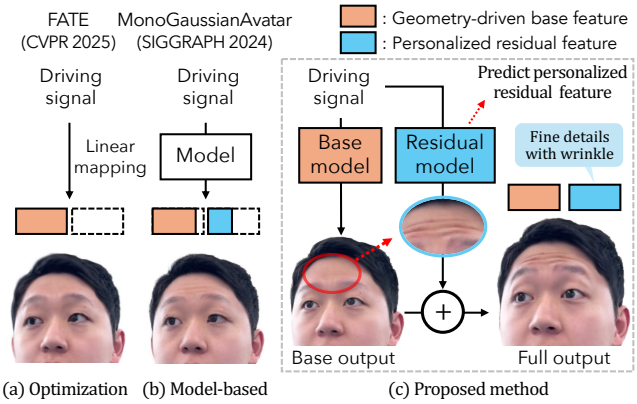

Figure 1: Conceptual comparison with prior approaches. (a) Optimization-based methods fail to capture residual details. (b) Entangled models suffer from learning ambiguity. (c) Our disentangled design separately models base and residual features for faithful reconstruction.

We identify the root cause of these limitations by rethinking the shared problem formulation that underlies these methods. As depicted in Fig. 1, we conceptualize facial features into two categories: *(i)* geometry-driven base features, which are well-represented by a 3DMM that captures the average facial variations across many individuals (*e.g.*, overall skin tone and rigid structural details); and *(ii)* residual features that capture high-frequency, person-specific details such as wrinkles and subtle skin deformations not represented within the expressive range of 3DMMs.

Analyzing prior works through this disentangled perspective reveals the source of their shortcomings. Despite their architectural differences, most 3D head creation methods share the common, challenging formulation of mapping a low-dimensional set of 3DMM parameters to the complex, high-dimensional space of facial attributes. Solving this problem with an optimization-based approach (Fig.1a, (Zhang et al. 2025)) can effectively capture stable base features via linear mappings from 3DMM parameters (Zhao et al. 2024; Qian et al. 2024; Shao et al. 2024; Zhang et al. 2025). However, their inherent linearity hinders model-

ing of non-linear residual details, even with subject-specific optimization. In contrast, more recent feature prediction models (Fig. 1b, (Chen et al. 2024)) attempt to address these limitations using a model-based approach that learns expression-dependent feature deformation from canonical space (Zielonka, Bolkart, and Thies 2023; Kirschstein et al. 2023; Xiang et al. 2024; Chen et al. 2024). However, this holistic approach forces the network to learn both base and residual features simultaneously, a complex task that results in an entangled representation where the crucial, personalized details are suppressed.

To resolve these limitations, we propose the **DipGuava** (Disentangled and Personalized Gaussian UV Avatar), a novel framework that rethinks the problem formulation. Our approach implements this as a functional disentanglement within a two-stage training, decomposing the 3D Gaussians' feature representation into simpler, more manageable subproblems as illustrated in Fig. 1c. Our two-stage pipeline is intentionally designed to facilitate this separation, even when supervised by the same ground truth images. We begin by learning a *geometry-driven base feature* in the first stage. By intentionally constraining the input of base network to local, first-order geometric information (unwrapped mesh surface normals), we guide it to learn a stable and low-frequency foundation that represents a robust average appearance. These features are predicted in UV-space and mapped onto the Gaussians bound on the driving mesh as their color and opacity attributes. This results in a coarse but stable representation, where each Gaussian carries base-level appearance aligned to the underlying geometry.

Once the stable base is established, a second network conditioned on both surface normals and FLAME parameters (serving as high-level descriptors) predicts the personalized residual features. These residuals, also defined in UV-space, are combined with the base feature by dynamic appearance fusion accounting for geometric deformations of each Gaussian. This separation and fusion enable the model to focus its capacity on capturing high-frequency, non-linear details (e.g., wrinkles) that the base model alone cannot represent.

Our contributions can be summarized as follows: 1. We propose DipGuava, a first 3DGS-based framework explicitly disentangling facial appearance into geometry-driven base and personalized residual, significantly reducing learning ambiguity. 2. We introduce a UV-based feature representation that enables dynamic appearance fusion, ensuring the effective integration of disentangled features from each stage. 3. Extensive evaluations demonstrate that DipGuava significantly outperforms existing approaches achieving superior expressiveness and photorealism.

2. Related Work

Creating 3D head avatars from monocular video remains a central challenge in vision and graphics. A widely used strategy builds upon 3DMMs (Paysan et al. 2009; Gerig et al. 2018; Li et al. 2017), which provide consistent mesh topology and interpretable control over shape and expression parameters. While effective for capturing coarse facial geometry, 3DMMs inherently lack the capacity to represent fine-grained details and non-linear skin deformations.

To overcome these limitations, recent works combine 3DMMs with neural rendering techniques, including SDFs (Park et al. 2019; Wang et al. 2021; Zheng et al. 2022a), triplanes (Xu et al. 2023; Ma et al. 2023), and NeRFs (Hong et al. 2022; Zhuang et al. 2022; Athar et al. 2022; Yao et al. 2022; Athar, Shu, and Samaras 2023). A common approach is to learn mappings from a canonical space (neutral expression) to posed and expressive faces. INSTA (Zielonka, Bolkart, and Thies 2023) leverages dynamic implicit fields with FLAME guidance, IMAvatar (Zheng et al. 2022b) uses blendshape-driven SDFs, and point-based methods like PointAvatar (Zheng et al. 2023) and GPAvatar (Chu et al. 2024) employ expression-conditioned point representations for detailed modeling.

The emergence of 3DGS (Kerbl et al. 2023) has enabled new approaches that use Gaussian points defined in a geometrically explicit form for rendering. Methods like GaussianAvatars (Qian et al. 2024) and SplattingAvatar (Shao et al. 2024) define Gaussians relative to mesh surfaces or points, animated via FLAME-driven LBS. Additionally, PSAvatar (Zhao et al. 2024) aims to represent components outside the face, such as glasses and hair, by geometrically assigning points. Gaussian blendshape models were also proposed (Ma et al. 2024; Li et al. 2025), where Gaussian features are conditioned on FLAME expression and pose parameters to capture facial dynamics. FATE (Zhang et al. 2025) bakes dynamic appearance into Gaussian texture maps, enabling stylization and unseen view regularization. Several recent works enhance generalization by leveraging Gaussian priors trained on diverse subjects. GEM (Zielonka et al. 2025a) builds a linear basis for lightweight distillation, while HeadGAP (Zheng et al. 2024), SynShot (Zielonka et al. 2025b), and SEGA (Guo et al. 2025) fine-tune subject-specific models for stable animation. While successful in capturing geometry and motion, these methods still limit their expressiveness to that of the driving 3DMMs.

To address this limited detail expressiveness, several methods have proposed modeling *dynamic* changes in Gaussian point features based on expression (Xiang et al. 2024; Xu et al. 2024; Chen et al. 2024). For instance, FlashAvatar (Xiang et al. 2024) defines geometric deformation in UV space to enhance training efficiency, while MonoGaussianAvatar (Chen et al. 2024) models expression-conditioned 3D Gaussian deformation fields for monocular reconstruction through per-Gaussian point feature representation. However, a key challenge remains in the complexity of mapping from the canonical space to the final expressed appearance, making it difficult for the model to disentangle geometry-driven cues from independently moving, identity-specific details such as subtle expressions and wrinkles.

In contrast, we propose **DipGuava**, a two-stage framework that *explicitly* disentangles facial appearance into geometry-driven base and personalized residual. Rather than jointly predicting both components, we first build a stable base feature and then separately predict personalized residuals. This separation reduces learning ambiguity and enables the network to capture fine-grained, identity-specific variations even under monocular supervision.

3. Preliminaries

In our work, a 3D head avatar is composed of multiple Gaussian primitives, each modeled as a spatially localized ellipsoid that captures local geometry and color. Each Gaussian comprises appearance features (color $c \in \mathbb{R}^3$ and opacity $o \in \mathbb{R}$) and geometric features (position $\mu \in \mathbb{R}^3$, scale $s \in \mathbb{R}^3$, and rotation $r \in \mathbb{R}^4$), which we refer to throughout this paper.

3.1 3D Gaussian Binding on Driving Mesh

Our approach animates facial expressions by adaptively deforming 3D Gaussians bound to the driving FLAME mesh using the binding mechanism in GaussianAvatars (Qian et al. 2024). This binding enables the Gaussians to first move with the facial mesh, capturing the linear movements defined by the 3DMM and to be deformed in local coordinates. The local space is defined by the triangle's centroid T, with the rotation matrix R capturing the triangle's orientation and a scaling factor k representing the mean edge length. Each Gaussian is initialized with position μ_l at the origin (each triangle's centroid), rotation r_l as an identity matrix, and scale s_l as a unit vector. These local geometric features are transformed into global space before rendering by:

$$r = Rr_l, \quad \mu = kR\mu_l + T, \quad s = ks_l. \tag{1}$$

3.2 Adaptive UV Feature Sampling

To establish a shared representation for feature prediction and fusion, we define Gaussian features in the UV space of the FLAME mesh. For Gaussians not lying directly on the mesh surface, we compute UV coordinates via barycentric interpolation over their associated triangles:

$$p_{\mu_l} = \mathbf{UV}(\mu_l), \quad \mathbf{x} = \text{Sample}(\mathcal{X}, p_{\mu_l}), \tag{2}$$

where $\mathcal{X}$ is the UV feature map and $\mathbf{x}$ is the sampled feature. Additional details are provided in the supplementary material.

4. Method

The overall pipeline of DipGuava is illustrated in Figure 2. To model fine, identity-specific dynamics beyond the expressive range of 3DMMs, we adopt a two-stage training scheme that explicitly disentangles geometry-driven base appearance from personalized residuals. This separation allows the model to first learn a stable, interpretable foundation, then refine it with residual details, enabling accurate modeling of non-linear deformations and subtle expressions.

4.1 Disentanglement of Base and Residual Features

Geometry-driven Base Appearance Training We begin by establishing a *geometry-driven base appearance* that represents the global facial structure and attributes aligned with the driving expression and pose. To achieve this, we rasterize the FLAME model's normals into UV space, producing a geometry-aware UV normal map $\mathcal{U}$. This is done via barycentric interpolation of vertex normals oriented with respect to the camera, allowing the UV map to accurately capture local surface orientation and curvature. This unwrapped normal map $\mathcal{U}$ is then fed into the **Base appearance network** $\mathcal{F}_{\text{base}}$, a U-Net architecture, which learns to predict the base appearance map $\mathcal{B}$.

During stage 1, for each 3D Gaussian primitive i, we jointly optimize its geometric features μ_l, s_l, r_l. To determine the base color and opacity for this Gaussian, we sample the base appearance map $\mathcal{B}$ at the corresponding UV coordinates p_{μ_l}, obtained by projecting the Gaussian's position onto the UV space: $p_{\mu_l} = \mathbf{UV}(\mu_l)$. The base color c_{b} and opacity o_{b} are then sampled as $(c_{\text{b}}, o_{\text{b}}) = \text{Sample}(\mathcal{B}, p_{\mu_l})$. Combined with the geometric attributes μ, s, r which are converted from their local forms μ_l, s_l, r_l into global space as described in Sec. 3.1, the full Gaussian representation at the first stage is:

$$G = \{c_{\text{b}}, o_{\text{b}}, \mu, s, r\}. \tag{3}$$

Personalized Residual Appearance Prediction The second stage focuses on modeling non-linear facial features, including high-frequency details like fine wrinkles and hair strands, which are not fully captured by the geometry-driven base feature. To this end, residual features for both appearance and geometry are predicted, aiming to reduce the discrepancy between the stage 1 output and the ground truth image.

With the geometric positions of the Gaussian points optimized in Stage 1 and the Base Appearance Network $\mathcal{F}_{\text{base}}$ frozen, we train a **Dynamic appearance network** $\mathcal{F}_{\text{d}}$. This network takes the same normal-based UV map $\mathcal{U}$ as input, along with a condition vector comprising FLAME expression parameters ψ and pose parameters θ. This dual-input design is intentional: while the normal map provides dense, low-level geometric detail, the FLAME parameters act as a compact, high-level semantic guide for the expression. This condition vector is concatenated at the U-Net's bottleneck to modulate the network's output toward the desired expression and pose. The network predicts a residual appearance map $\mathcal{R}$ as:

$$\mathcal{R} = \mathcal{F}_{\text{d}}(\mathcal{U}, \psi, \theta). \tag{4}$$

In parallel, we employ a geometric MLP, $\mathcal{F}_{\text{g}}$, which takes the same condition vector (FLAME expression parameters ψ and pose parameters θ) as input and predicts a geometric deformation map $\Delta\mathcal{G}$ in UV space:

$$\Delta\mathcal{G} = \mathcal{F}_{\text{g}}(\psi, \theta). \tag{5}$$

This design enables the prediction of both appearance and geometric residuals conditioned on shared FLAME-based semantic parameters, thus maintaining a clear disentanglement from the base geometry-driven representation established in the first stage.

4.2 Dynamic Appearance Fusion

Simply adding residual features sampled from the original location of a Gaussian point fails to account for deformation-induced shifts. As a result, these residual features can become misaligned with the intended facial region, leading to appearance artifacts or blurred details. To address this, we introduce a geometry-aware residual fusion strategy that resamples dynamic appearance features at the updated UV coordinates *after* the geometric deformation. This ensures that

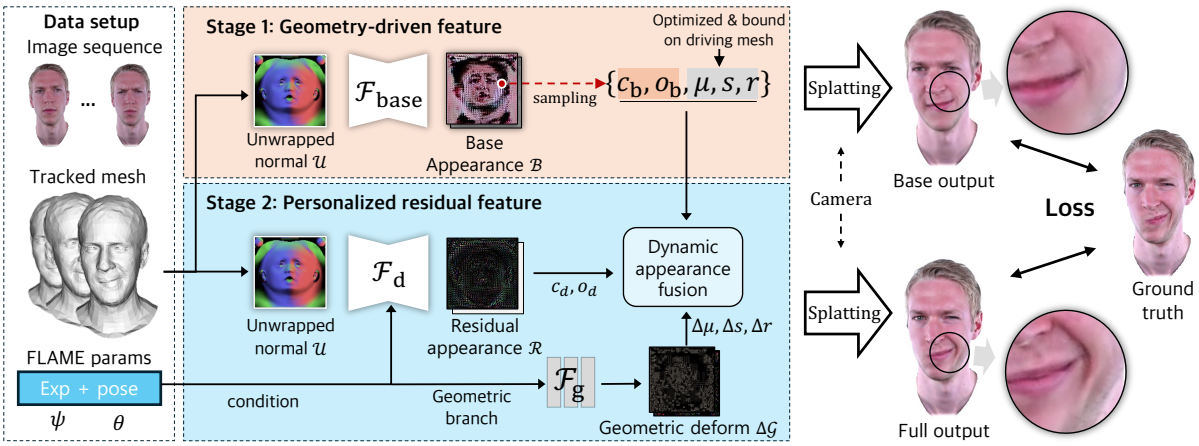

Figure 2: **Overview of DipGuava**. Stage 1 optimizes a geometry-driven base appearance (overall color and opacity from mesh surface normals). Stage 2 predicts residual features to capture facial details beyond the base appearance. Dynamic appearance fusion combines these residuals with the base appearance and geometric deformations.

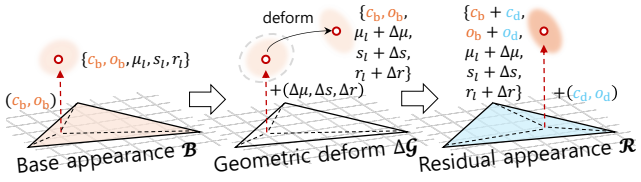

Figure 3: **Dynamic appearance fusion.** To ensure alignment, personalized residual appearance sampled from UV space is combined with the base appearance *after* the geometric deformation.

residual appearance features are sampled from the UV location corresponding to the deformed Gaussian's semantic context. This design avoids erroneous blending and allows for more precise correction of expression-specific variations.

During the fusion process, the Gaussian features are obtained by combining geometry-driven and residual features, as depicted in Figure 3. Each Gaussian's optimized position μ_l from the first stage is used to sample its base appearance (c_b, o_b). Next, residual geometric deltas in local coordinates $(\Delta\mu, \Delta s, \Delta r)$ are sampled from $\Delta\mathcal{G}$ and applied as:

$$\mu_l' = \mu_l + \Delta\mu, \quad s_l' = s_l + \Delta s, \quad r_l' = r_l + \Delta r. \tag{6}$$

At the updated UV position, personalized residuals $(c_\mathrm{d}, o_\mathrm{d})$ are then sampled from $\mathcal{R}$ and added to the base:

$$p_{\mu_l'} = \mathbf{UV}(\mu_l'), \quad (c_\mathrm{d}, o_\mathrm{d}) = \mathrm{Sample}(\mathcal{R}, p_{\mu_l'}), \tag{7}$$

$$c = c_\mathrm{b} + c_\mathrm{d}, \quad o = o_\mathrm{b} + o_\mathrm{d}. \tag{8}$$

After transforming the geometry into global space, the final representation becomes:

$$G = \{c, o, \mu', s', r'\}. \tag{9}$$

4.3 Gaussian Splatting Rendering

We render the output image using 3D Gaussian splatting as introduced by Kerbl *et al.* (Kerbl et al. 2023). In our framework, as shown in Fig. 2, this rendering process is applied at both training stages. During stage 1, only the geometry-driven base features are used for rendering. In stage 2, we render the fused features that incorporate residual appearance and residual geometric deformation.

4.4 Training Objectives

Both Stage 1 and Stage 2 are supervised with the same GT images. In Stage 1, the loss is computed from the rendered output using geometry-driven base features. In Stage 2, it is computed on the final fused image after dynamic appearance fusion. The total loss includes photometric, perceptual, and geometric regularization terms.

$$\mathcal{L}_\mathrm{total} = \mathcal{L}_\mathrm{pho} + \lambda_\mathrm{lpips}\mathcal{L}_\mathrm{lpips} + \lambda_\mathrm{xyz}\mathcal{L}_\mathrm{xyz} + \lambda_\mathrm{scale}\mathcal{L}_\mathrm{scale}. \tag{10}$$

We use a combination of L_1 and SSIM losses:

$$\mathcal{L}_\mathrm{pho} = \lambda_\mathrm{L1}|I - I_\mathrm{gt}|_1 + (1 - \lambda_\mathrm{L1})\,(1 - \mathrm{SSIM}(I, I_\mathrm{gt}))\,. \tag{11}$$

To encourage perceptual similarity, we include an LPIPS loss (Zhang et al. 2018):

$$\mathcal{L}_\mathrm{lpips} = |\phi(I) - \phi(I_\mathrm{gt})|_2^2. \tag{12}$$

where ϕ denotes a VGG-based feature extractor. Following prior work (Qian et al. 2024), we regularize both the original and deformed 3D Gaussian primitives by penalizing out-of-bounds positions and overly large scales with thresholds ϵ_xyz and ϵ_scale. Specifically, the loss is applied to both the stage-1 geometry (μ_l, s_l) and the predicted stage-2 geometry (μ_l', s_l').

$$\mathcal{L}_\mathrm{xyz} = \sum_{\tilde{\mu}\in\{\mu_l,\mu_l'\}} \left\|\max\left(\|\tilde{\mu}\|_2 - \epsilon_\mathrm{xyz},\ 0\right)\right\|_2, \tag{13}$$

$$\mathcal{L}_\mathrm{scale} = \sum_{\tilde{s}\in\{s_l,s_l'\}} \left\|\max\left(\exp(\tilde{s}) - \epsilon_\mathrm{scale},\ 0\right)\right\|_2. \tag{14}$$

Method	L1 ↓	LPIPS ↓	SSIM ↑	PSNR ↑
PointAvatar	0.016	0.056	0.923	25.21
INSTA	0.016	0.108	0.907	24.73
GaussianAvatars	<u>0.012</u>	0.066	0.944	28.36
FlashAvatar	0.018	0.071	0.918	25.72
SplattingAvatar	0.019	0.096	0.917	25.61
MonoGaussianAvatar	0.014	0.053	0.939	26.37
FATE	0.012	<u>0.045</u>	<u>0.945</u>	<u>28.38</u>
Ours	**0.009**	**0.042**	**0.957**	**30.14**

Table 1: Quantitative comparison with SOTA methods. **Best** and <u>second-best</u> results highlighted.

5. Experiments

Datasets Our method uses a monocular video of a single subject as input. We use videos from NHA (Grassal et al. 2022), NerFace (Gafni et al. 2021), PointAvatar (Zheng et al. 2023), and INSTA (Zielonka, Bolkart, and Thies 2023) for 11 subjects. Additionally, we captured 7 more videos under diverse conditions, including both controlled studio settings and uncontrolled scenarios using a mobile phone. In total, 18 videos with various environments were used in our experiments. Each video ranges 1–3 minutes (512x512 resolution). For a fair comparison, we use the same tracking results (FLAME parameters and meshes with teeth and vertex deformations around the hairline contour), masked images (Yu et al. 2021), face-camera rotation from VHAP (Qian 2024) across all sequences, and we set the last 30% of frames from each subject as the testing set.

Implementation Details We train the model in two stages, each for 60k iterations. In stage 1, we jointly optimize the base geometry and $\mathcal{F}_{\text{base}}$ using a learning rate of 1×10^{-4}. Densification, cloning, and splitting are performed every 500 iterations, with Gaussians having opacity below 0.05 pruned. After stage 1, both the optimized geometry and the weights of $\mathcal{F}_{\text{base}}$ are frozen. In stage 2, the residual appearance network and geometric MLP are trained with a learning rate of 1×10^{-5}. We use the following loss weights: $\lambda_{\text{L1}} = 0.8, \lambda_{\text{LPIPS}} = 0.1, \lambda_{\text{xyz}} = 0.01, \lambda_{\text{scaling}} = 1.0$, and geometric threshold in local scale: $\epsilon_{\text{xyz}} = 2.0, \epsilon_{\text{scale}} = 0.6$. All UV maps have a resolution of 128x128.

5.1 Comparison with SOTA Methods

To evaluate the effectiveness of proposed method, we conducted comprehensive experiments comparing DipGuava with a wide range of SOTA 3D head avatar approaches. These include PointAvatar (PA) (Zheng et al. 2023), INSTA (Zielonka, Bolkart, and Thies 2023), GaussianAvatars (GA) (Qian et al. 2024), FlashAvatar (FA) (Xiang et al. 2024), SplattingAvatar (SA) (Shao et al. 2024), MonoGaussianAvatar (MGA) (Chen et al. 2024), and FATE (Zhang et al. 2025). Among these, we include qualitative comparisons with methods that produce reasonably comparable outputs. Please refer to the supplementary material for visual comparisons against all methods and per-subject results.

Training Time and Inference Speed Under identical settings (RTX A6000), our method converges in 70 minutes, which is comparable to existing approaches (FATE: 45m, MGA: 9h, SA: 44m, FA: 20m, GA: 40m, INSTA: 60m, PA: 7h), while achieving superior performance in capturing subtle expressions and details. At inference time, our full model runs at 88 FPS (512×512 resolution), enabling real-time applications.

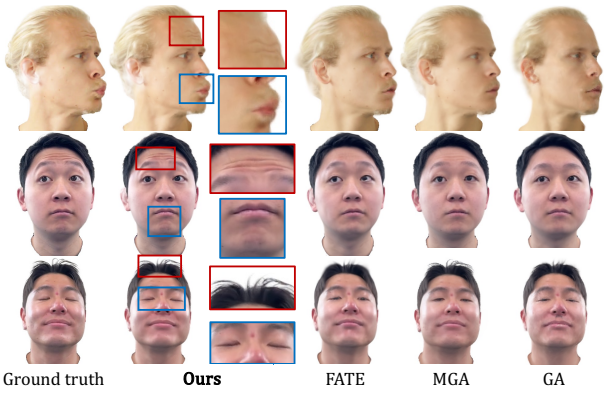

Figure 4: **Qualitative comparison in self-driven animation.** Our method generates outperforming results with details such as wrinkles, eye blinks, and lip movement.

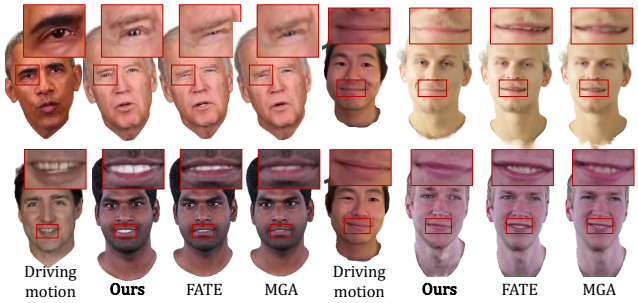

Figure 5: **Qualitative comparison in cross-id reenactment.** The proposed method preserves both facial structure and appearance with high fidelity, while accurately following subtle expressions in the driving motion.

Quantitative Performance Comparison For quantitative evaluation, we use L1 distance, LPIPS (Zhang et al. 2018), SSIM (Wang et al. 2004), and PSNR. Table 1 summarizes the average performance across all identities in our benchmark. DipGuava consistently outperforms all baselines across all metrics, clearly demonstrating its effectiveness under identical training conditions. The performance gap arises from the fact that only our method accurately reconstructs fine-grained, identity-specific attributes. While previous approaches may look acceptable without ground truth, they fail to capture the personalized details critical for realism. This distinction is evident in the following qualitative results.

Self-driven Re-animation Figure 4 presents a qualitative comparison of DipGuava with prior state-of-the-art methods. Overall, DipGuava produces sharper and more vivid outputs while faithfully reconstructing personalized appear-

ance and expression-driven geometry.

In both the first and second subjects, DipGuava is the only method that accurately reconstructs *forehead wrinkles*, which are entirely missing or oversmoothed in other methods. This highlights DipGuava's strength in preserving identity-specific high-frequency features. For the first subject, DipGuava captures subtle lip protrusion under a side-profile view, which others fail to represent. In the second subject, the jawline and eyelids follow expression-induced shape changes with clear articulation, especially in wide-eye expressions with upward gaze. The third subject further demonstrates DipGuava's robustness in reconstructing closed-eye expressions without introducing artifacts, unlike prior works that either hallucinate eyes or leave them open. Additionally, fine structures such as raised bangs are reconstructed with greater accuracy.

Cross-identity Reenactment In Fig. 5, we evaluate cross-identity reenactment performance, comparing our method with recent models which show reasonable results in this task. A key distinction lies in the fidelity of expression transfer. DipGuava more accurately conveys subtle motions, such as slight smiles and eye blinks, while preserving identity. In contrast, both FATE and MGA often exhibit entangled modeling of lips and teeth, frequently revealing teeth even in closed-mouth expressions, and rendering eye movements unnaturally. Notably, DipGuava consistently reproduces gaze direction and eyelid motion with higher precision, even in unseen, subtle expressions such as closed-mouth smiles or eyelid behavior.

5.2 Analysis of Training and Model Design Choices

Since our method integrates a two-stage training pipeline with multiple components, we conduct a comprehensive analysis to validate the effectiveness of each design choice as shown in Tab. 2.

Effect of Training Strategy We compare two variants to evaluate convergence and detail fidelity. ***Joint*** fine-tunes both $\mathcal{F}$base and residual features with a reduced learning rate for the base. ***LowRes*** trains $\mathcal{F}$base at 256^2 resolution, then switches to 512^2 for residual refinement. Our staged strategy achieves superior convergence and recovers high-frequency details more effectively.

Effect of Model Design Choice We assess key architectural components via ablations. ***OptBase*** removes surface normal input and optimizes appearances, degrading generalization to unseen expressions. ***FixedUV*** disables adaptive UV sampling, limiting expression-specific details. ***MLP*** replaces the UNet backbone with an MLP-based model, resulting in worse performance, especially in high-frequency regions. This confirms the benefit of spatially-aware convolution in UV space.

5.3 Ablations for Model Components

We conduct detailed ablations to evaluate the contribution of each component in our framework as shown in Tab. 2 and Fig. 6. Here, "***only*** $\mathcal{B}$" and "***only*** $\mathcal{R}$" represent single-stage models that predict appearance using only the base or residual branch, respectively. For the two-stage setup, we evaluate "***w/o*** $\mathcal{R}$", "***w/o*** $\Delta\mathcal{G}$", and disabling dynamic appearance fusion by simply summing $\mathcal{B}$ and $\mathcal{R}$ ("***w/o DAF***").

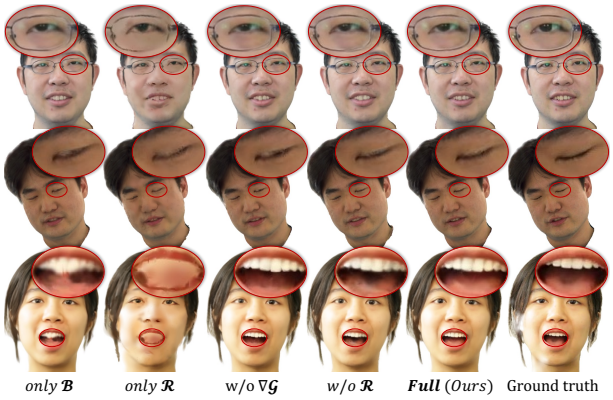

Figure 6: **Impact of individual components.** The full model preserves structure and fine-scale details most effectively.

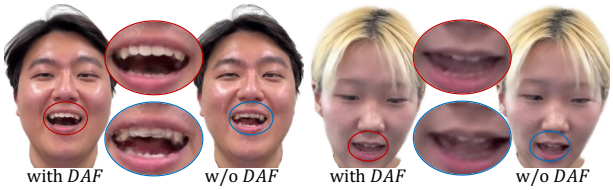

Figure 7: Applying **dynamic appearance fusion** yields sharper and less noisy results in regions with rapid color changes, such as inside the mouth.

Effectiveness of Two-stage Feature Disentanglement Fig. 6 shows that while the base appearance captures the overall facial layout, personalized wrinkles and high-frequency attributes such as eyeglass frames are preserved only when residual features are incorporated. Relying solely on $\mathcal{R}$ leads to significant performance drops across all metrics and introduces noticeable artifacts like blurred fine structures that 3DMMs struggle to represent. Our full two-stage model consistently outperforms all ablations, demonstrating the benefit of explicit feature disentangling.

Complementary Roles of Residual Appearance and Geometric Deformation Within our two-stage framework, the residual appearance and geometric deformation play distinct yet complementary roles. As shown in Table 2 and Fig. 6, removing either residual component leads to a consistent drop in performance, indicating that relying on either color correction or geometric adjustment alone is insufficient to capture subtle motion and appearance. For example, accurate eyeglass reconstruction requires both color refinement from the residual appearance and precise geometric adjustments to align the Gaussians with the frame's shape.

Importance of Dynamic Appearance Fusion Interestingly, while *w/o DAF* achieves comparable performance in LPIPS compared to proposed full method, Fig. 7 reveals that

	Training strategy		Model design			Model component					***Full***
Metric	*Joint*	*LowRes*	*OptBase*	*FixedUV*	*MLP*	*only* $\mathcal{B}$	*only* $\mathcal{R}$	*w/o* $\Delta\mathcal{G}$	*w/o* $\mathcal{R}$	*w/o DAF*	$\mathcal{B}$+$\mathcal{R}$+$\Delta\mathcal{G}$
L1 ↓	0.013	0.011	0.012	0.010	0.010	0.0099	0.0130	0.0093	0.0092	<u>0.0091</u>	**0.0090**
LPIPS ↓	0.0491	0.0949	0.0492	0.0453	0.0516	0.0433	0.0955	0.0440	0.0434	**0.0410**	<u>0.0417</u>
SSIM ↑	0.936	0.949	0.944	0.952	0.938	0.9538	0.9357	0.9547	0.9553	<u>0.9554</u>	**0.9570**
PSNR ↑	29.02	29.42	28.70	29.53	28.32	29.54	28.25	29.94	<u>29.95</u>	29.86	**30.14**

Table 2: Quantitative result of training strategies, model designs, and ablations. **Best** and <u>second-best</u> results are highlighted.

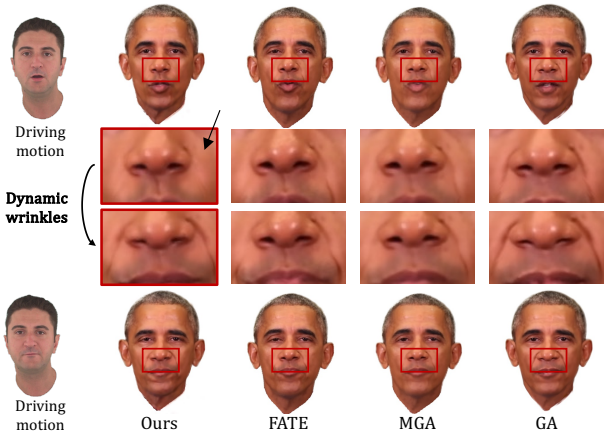

Figure 8: **Expression-aware dynamic wrinkles.** Unlike other methods that model wrinkles as static color textures, our method adaptively captures them in response to facial motion and geometry.

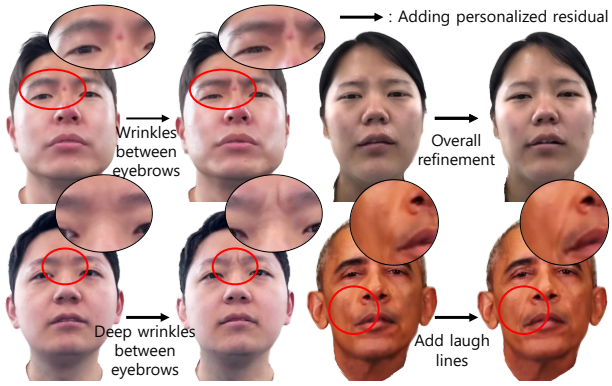

Figure 9: **Personalized residuals from the same driving motion.** Residual features inject subject-specific details, enabling personalized expressions from identical motions.

omitting dynamic appearance fusion leads to noise and artifacts in semantically complex regions like the mouth interior. In other words, applying residual appearance features from pre-deformation UV coordinates results in mismatched textures. In contrast, our geometry-aware fusion re-samples residuals *after* deformation, ensuring proper alignment and enabling the model to capture non-linear variations more effectively.

5.4 Analysis of Personalized Residual Feature

Dynamic Wrinkle Modeling Fig. 8 further illustrates the dynamic wrinkle modeling ability of DipGuava. Unlike other methods that render wrinkles as static textures unaffected by expression, our method generates expression-dependent wrinkles in a natural and temporally coherent manner. For example, in the upper row, wrinkles around the mouth are smoothed as the lips move forward, while in the lower row, new wrinkles emerge in response to the expression, faithfully reflecting the underlying skin deformation.

Id-specific Attribute in Personalized Residuals Beyond simply enhancing detail, our method models personalized residuals in an identity-specific manner, as shown in Fig. 9. While the base model captures the shared expression across subjects, the residual features add unique, subject-specific attributes, effectively disentangling individual characteristic from the shared motion. For example, the model accurately synthesizes localized wrinkles for subjects who naturally exhibit them, while preserving smooth skin for others without introducing hallucinated details. This demonstrates that our method learns not only to refine appearance, but also to model personalized, identity-consistent dynamics.

6. Conclusion and Limitations

We presented DipGuava, a novel framework that creates high-fidelity, animatable 3D head avatars from monocular video. Our key contribution is a two-stage approach that explicitly disentangles the complex facial representation into a geometry-driven base and a personalized residual. This decomposition resolves the learning ambiguity of prior methods, enabling the targeted modeling of high-frequency, subject-specific details. Finally, combined with a dynamic fusion mechanism, our method achieves fine-grained reconstruction of personalized facial features. Extensive experiments across multiple benchmarks demonstrate that DipGuava outperforms previous methods, offering a practical solution for real-world 3D head avatar applications. However, our method shares a inherent limitation among 3DMM-driven approaches such as contrained robustness by the fidelity of the 3DMM tracking and the model's capacity to represent poorly sampled areas, such as intra-oral regions or extreme expressions. Nevertheless, we argue that our disentangled architecture is highly scalable. As tracking fidelity improves, our base model provides a cleaner foundation, enabling the residual model to learn finer details and further widen the performance gap over holistic approaches.

Acknowledgements

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (RS-2024-00398413, Contribution Rate: 90%), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-02216328), and the Yonsei Signature Research Cluster Program of 2025 (2025-22-0013).

References

Athar, S.; Shu, Z.; and Samaras, D. 2023. Flame-in-nerf: Neural control of radiance fields for free view face animation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–8. IEEE.

Athar, S.; Xu, Z.; Sunkavalli, K.; Shechtman, E.; and Shu, Z. 2022. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 20364–20373.

Chen, Y.; Wang, L.; Li, Q.; Xiao, H.; Zhang, S.; Yao, H.; and Liu, Y. 2024. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, 1–9.

Chu, X.; Li, Y.; Zeng, A.; Yang, T.; Lin, L.; Liu, Y.; and Harada, T. 2024. GPAvatar: Generalizable and precise head avatar from image (s). *arXiv preprint arXiv:2401.10215*.

Gafni, G.; Thies, J.; Zollhofer, M.; and Nießner, M. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8649–8658.

Gerig, T.; Morel-Forster, A.; Blumer, C.; Egger, B.; Luthi, M.; Schönborn, S.; and Vetter, T. 2018. Morphable face models-an open framework. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 75–82. IEEE.

Grassal, P.-W.; Prinzler, M.; Leistner, T.; Rother, C.; Nießner, M.; and Thies, J. 2022. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18653–18664.

Guo, C.; Su, Z.; Wang, J.; Li, S.; Chang, X.; Li, Z.; Zhao, Y.; Wang, G.; and Huang, R. 2025. SEGA: Drivable 3D Gaussian Head Avatar from a Single Image. *arXiv preprint arXiv:2504.14373*.

Hong, Y.; Peng, B.; Xiao, H.; Liu, L.; and Zhang, J. 2022. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20374–20384.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kirschstein, T.; Qian, S.; Giebenhain, S.; Walter, T.; and Nießner, M. 2023. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4): 1–14.

Li, L.; Li, Y.; Weng, Y.; Zheng, Y.; and Zhou, K. 2025. RGBAvatar: Reduced Gaussian Blendshapes for Online Modeling of Head Avatars. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10747–10757.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6): 194–1.

Ma, S.; Weng, Y.; Shao, T.; and Zhou, K. 2024. 3d gaussian blendshapes for head avatar animation. In *ACM SIGGRAPH 2024 Conference Papers*, 1–10.

Ma, Z.; Zhu, X.; Qi, G.-J.; Lei, Z.; and Zhang, L. 2023. Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16901–16910.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 165–174.

Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; and Vetter, T. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, 296–301. Ieee.

Qian, S. 2024. VHAP: Versatile Head Alignment with Adaptive Appearance Priors. https://github.com/ShenhanQian/VHAP.

Qian, S.; Kirschstein, T.; Schoneveld, L.; Davoli, D.; Giebenhain, S.; and Nießner, M. 2024. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20299–20309.

Shao, Z.; Wang, Z.; Li, Z.; Wang, D.; Lin, X.; Zhang, Y.; Fan, M.; and Wang, Z. 2024. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1606–1616.

Wang, L.; Chen, Z.; Yu, T.; Ma, C.; Li, L.; and Liu, Y. 2022. FaceVerse: a Fine-grained and Detail-controllable 3D Face Morphable Model from a Hybrid Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2022)*.

Wang, X.; Guo, Y.; Yang, Z.; and Zhang, J. 2021. Prior-guided multi-view 3d head reconstruction. *IEEE Transactions on Multimedia*, 24: 4028–4040.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xiang, J.; Gao, X.; Guo, Y.; and Zhang, J. 2024. FlashAvatar: High-fidelity Head Avatar with Efficient Gaussian

Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1802–1812.

Xu, Y.; Chen, B.; Li, Z.; Zhang, H.; Wang, L.; Zheng, Z.; and Liu, Y. 2024. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1931–1941.

Xu, Y.; Wang, L.; Zhao, X.; Zhang, H.; and Liu, Y. 2023. Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–10.

Yao, S.; Zhong, R.; Yan, Y.; Zhai, G.; and Yang, X. 2022. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*.

Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; and Sang, N. 2021. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International journal of computer vision*, 129: 3051–3068.

Zhang, J.; Wu, Z.; Liang, Z.; Gong, Y.; Hu, D.; Yao, Y.; Cao, X.; and Zhu, H. 2025. Fate: Full-head gaussian avatar with textural editing from monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5535–5545.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhao, Z.; Bao, Z.; Li, Q.; Qiu, G.; and Liu, K. 2024. Psavatar: A point-based morphable shape model for real-time head avatar creation with 3d gaussian splatting. *arXiv preprint arXiv:2401.12900*.

Zheng, X.; Liu, Y.; Wang, P.; and Tong, X. 2022a. SDF-StyleGAN: implicit SDF-based StyleGAN for 3D shape generation. In *Computer Graphics Forum*, volume 41, 52–63. Wiley Online Library.

Zheng, X.; Wen, C.; Li, Z.; Zhang, W.; Su, Z.; Chang, X.; Zhao, Y.; Lv, Z.; Zhang, X.; Zhang, Y.; et al. 2024. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. *arXiv preprint arXiv:2408.06019*.

Zheng, Y.; Abrevaya, V. F.; Bühler, M. C.; Chen, X.; Black, M. J.; and Hilliges, O. 2022b. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13545–13555.

Zheng, Y.; Yifan, W.; Wetzstein, G.; Black, M. J.; and Hilliges, O. 2023. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21057–21067.

Zhuang, Y.; Zhu, H.; Sun, X.; and Cao, X. 2022. Mofanerf: Morphable facial neural radiance field. In *European conference on computer vision*, 268–285. Springer.

Zielonka, W.; Bolkart, T.; Beeler, T.; and Thies, J. 2025a. Gaussian eigen models for human heads. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15930–15940.

Zielonka, W.; Bolkart, T.; and Thies, J. 2023. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4574–4584.

Zielonka, W.; Garbin, S. J.; Lattas, A.; Kopanas, G.; Gotardo, P.; Beeler, T.; Thies, J.; and Bolkart, T. 2025b. Synthetic prior for few-shot drivable head avatar inversion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10735–10746.