개인화된 3차원 생성 모델을 활용한 고품질 발화 영상 생성

□ 고재훈, 김승룡 / 고려대학교

요 약

오디오 기반 발화 영상 생성을 위한 최근의 방법은 종종 단안 비디오에서 신경 방사장(NeRF)을 최적화하며, 고화질 및 3D 일관성 있는 새로운 뷰 프레임을 렌더링하는 기능을 활용한다. 그러나 단안 비디오에 포괄적인 3D 정보가 없기 때문에 완전한 얼굴 기하학을 reconstruct하는 데 어려움을 겪는 경우가 많다. 본 논문에서는 개인화를 통해 사전 훈련된 3차원 생성 모델을 바탕으로 그럴듯한 영상을 생성할 수 있는 새로운 오디오 기반 발화 영상 생성 프레임워크인 Talk3D를 제안한다. 본 모델은 개인화된 3D 생성 모델과 더불어 입력 오디오에 의해 구동되는 NeRF 공간의 동적 얼굴 변화를 예측하는 새로운 오디오 유도 주의 U-Net 아키텍처를 제시하며, 오디오와 무관한 장면 변화를 효과적으로 분리하는 다양한 컨디션 토큰을 사용하는 구조이다. 기존 방법에 비해 Talk3D는 극단적인 카메라 포즈에서도 현실적인 영상을 생성하는 데 탁월하며, 우리의 접근 방식이 정량적 및 정성적 평가 측면에서 최첨단 벤치마크를 능가한다는 것을 보여주는 광범위한 실험을 수행하고 이를 설명하고자 한다.

Ⅰ. 서 론

오디오 기반의 발화 영상 생성은 입력 오디오에 동기화된 입술 움직임이 있는 얼굴 영상을 생성하는 것을 목표로 한다. 이 작업은 음성을 정확하게 분석하고, 얼굴의 사실적인 움직임으로 바꾸며, 동시에 고품질의 이미지를 생성하는 다양한 과제를 포함한다. 초기 접근 방식은 이미지 reconstruction에 초점을 맞춰 2D 생성 모델을 사용하였지만, 이러한 방법은 종종 카메라 포즈 제어에 한계를 보인다. 이를 해결하기 위해 최근의 연구는 신경 방사장[9]을 활용하여 다중 뷰에서 일관적인 이미지를 렌더링하고 동시에 카메라 포즈 제어 가능성을 제공하고 있다. 그럼에도 불구하고, 단안 비디오에서 NeRF를 최적화하는 것은 여전히 어려운 문제이다. 이는 다양한 카메라 포즈와 3D 정보의 부족에서 비롯되며, 단안 비디오 데이터에 존재하지 않는 카메라 방향에서 렌더링 품질이 부족해지게 된다. 최신 방법조차도 극단적인 카메라 포즈에서 고품질 이미지를 생성하는 데 어려움을 겪고 있다. (〈그림 1〉 참고)

이러한 문제를 해결하기 위해, 우리는 극단적인 카메

<그림 1>

라 포즈에서도 발화 영상을 생성하기 위해 고안된 새로운 프레임워크인 Talk3D를 제안한다. 우리의 방법의 핵심은 다양한 객체와 장면에서 현실적인 3D geometry를 합성할 수 있는 3D GAN[1,2,3]을 활용하는 것이다. 이 모델들은 적대적 훈련 전략을 3D 표현과 결합하여 2D 이미지 컬렉션에서 3D 기하학적 정보를 학습한다[1,2,3]. 그중에서도 특히, 다양한 연구들이 3D GAN과 GAN inversion 방법을 결합하여 이 기능을 활용하고자 시도하였고, 단일 얼굴 이미지에서 3차원 재구성에 성공하였다[4,5,6]. 그러나 이러한 접근법들은 오디오 기반의 발화 문제를 다루지 않고 있다. 우리는 3D GAN의 특성이 3D 발화 영상 생성 영역으로도 확장될 수 있다고 제안한다. 이러한 모델을 구현하기 위한 직관적인 전략 중 하나는 주어진 오디오 입력을 사용하여 GAN 잠재 공간 내에서 잠재 벡터를 직접 예측하는 것으로, 과거 연구 중 이 전략을 사용하여 GAN 잠재 공간에서 오디오 기반 편집을 수행하였다[7]. 그러나 이 접근법은 high-dimensional한 잠재 벡터가 입술 영역뿐만 아니라 얼굴의 personal identity와 배경을 포함한 전체 장면 구조를 나타내는 등의 feature entanglement 문제로 어려움을 겪게 된다. 발화 영상 생성에서는 다른 영역을 고정하면서 입술이나 눈과 같은 국부적인 영역만 편집해야 하기 때문이다. 3D GAN을 효과적으로 통합하기 위해, 우리의 Talk3D는 GAN 잠재 공간 대신 3차원 공간인 NeRF의 직접적인 편집을 수행하여 주어진 오디오 입력에 정확히 입술을 동기화하고, 현실적인 얼굴 아바타를 생성한다. 구체적으로, 우리의 모델은 3D GAN을 특정 personal identity에 미세 조정하는 간단하지만 효과적인 전략인 VIVE3D[8]를 채택한다. 그 이후에, 우리의 U-Net 기반 아키텍처는 미세 조정된 triplane을 사용하여 triplane 오프셋, 즉 deltaplane을 생성한다. 오디오에 의해 우리의 모델은 해당 오디오에 맞춰 입술 움직임을 정확히 나타내는 deltaplane을 예측하게 된다. 우리의 U-Net 아키텍처는 attention 기반 모듈을 채택하고 다양한 조건부 토큰으로 추가 feature를 적용하여 입술 움직임에서 미세한 변화를 분리(예: 상체, 배경, 눈 움직임)하여 이미지 재구성 품질과 입술 동기화 정확도를 향상시킨다. 광범위한 정성적 및 정량적 실험을 통해, 우리의 방법은 렌더링된

이미지뿐 아니라 정면 관점에서도 이전 접근법보다 생성 품질이 뛰어남을 보여준다.

Ⅱ. 개인화된 오디오 기반 NeRF 기술

1. 문제 정의

기존의 NeRF[9,10,11]는 단일 정적 장면에 최적화되는 것을 목표로 하지만, NeRF 기반의 3D GAN[1,2,3]은 NeRF 공간을 랜덤으로 샘플링된 잠재 코드 w로 조건화하여 명시적으로 포즈 제어 가능한 이미지 생성을 달성했다. 이 중에서 EG3D[1]는 세 가지 단계를 통해 우수한 성능을 보여준다. 먼저, EG3D는 평면 생성기 G를 사용하여 낮은 해상도의 특징 평면 P를 효율적으로 합성하며, 이는 $P=G(w; \theta_G)$로 표현된다. 이 특징 평면은 세 개의 직교하는 특징 평면 $\{P_{xy}, P_{yx}, P_{zx}\}$로 변형된다. EG3D는 이러한 직교 평면에서 concatenate된 특징을 받아서 밀도 σ와 feature f로 매핑하는 MLP를 이용한다. 이 특징 필드는 저해상도 2D 특징 맵 F로 렌더링된다. 마지막으로, 생성된 특징 맵 F는 여러 CNN layer로 구성된 2D super-resolution 모듈에서 처리되어 최종 이미지 I를 생성한다. 이 연속적인 과정(볼륨 렌더링 및 super-resolution 모듈 포함)을 $R(\cdot;\theta_R)$로 표기한다. 최종 합성 이미지는 다음과 같이 공식화된다: $I=R(P, \pi; \theta_R)$. 이때 π는 카메라 파라미터를 나타낸다.

이 절에서는 포즈 제어 가능한 오디오 기반 고품질 말하는 초상화 합성을 가능하게 하는 우리의 방법, Talk3D의 주요 구성 요소를 설명한다. 특정 personal identity에 대한 N개의 비디오 프레임 $V= I_n$가 주어졌을 때, 우리의 모델은 n번째 프레임 이미지 I_n와 해당 오디오 특징 a_n, 그리고 카메라 파라미터 π_n를 입력으로 받는다. 우리는 오디오 기반 렌더링 과정을 다음과 같이 공식화한다:

$$\begin{aligned} \mathbf{P} &= \mathcal{G}(\mathbf{w}; \theta_{\mathcal{G}}), \\ I'_n &= \mathcal{R}(\mathbf{P}, \pi_n, \mathbf{a}_n; \theta_{\mathcal{R}}). \end{aligned} \tag{1}$$

프레임 I_n의 입술 움직임을 가장 잘 재현하는 렌더링된 초상 이미지 I_n'를 얻기 위해, 우리 모델은 EG3D 파라미터 θ와 해당 프레임을 나타내는 최적의 triplane P′를 찾는 것을 목표로 한다. 추론 시, 새로운 오디오 a_n^{novel}가 주어졌을 때, 우리는 이것을 다음과 같이 재공식화 한다:

$$\begin{aligned} \mathbf{P}_{\mathrm{ID}} &= \mathcal{G}(\mathbf{w}_{\mathrm{ID}}; \theta^*_{\mathcal{G}}), \\ I_n^{\mathrm{novel}} &= \mathcal{R}(\mathbf{P}_{\mathrm{ID}}, \pi_n, \mathbf{a}_n^{\mathrm{novel}}; \theta^*_{\mathcal{R}}), \end{aligned} \tag{2}$$

여기서 w_{ID}는 특정 사람의 얼굴 정체성에 해당하는 identity 잠재 코드이다. 그 다음, 이러한 개인화된 생성기는 identity triplane인 P_{ID}를 생성한다. 이를 공식화하기 위해, 우리는 먼저 w_{ID}와 θ_G, θ_R를 제공하는 개인화된 3D GAN을 미세 조정한다. 그 이후 렌더러 R에서 a_n^{novel}를 조건화하기 위해, 우리는 a_n^{novel}로부터 새로운 평면 ΔP_n^{novel}을 생성하는 deltaplane 생성기를 제안하여 identity 평면 P_{ID}를 수정한다. 즉, $P'= P_{ID}+\Delta P_n^{novel}$로 하여 최종 렌더러는 다음과 같이 정의된다:

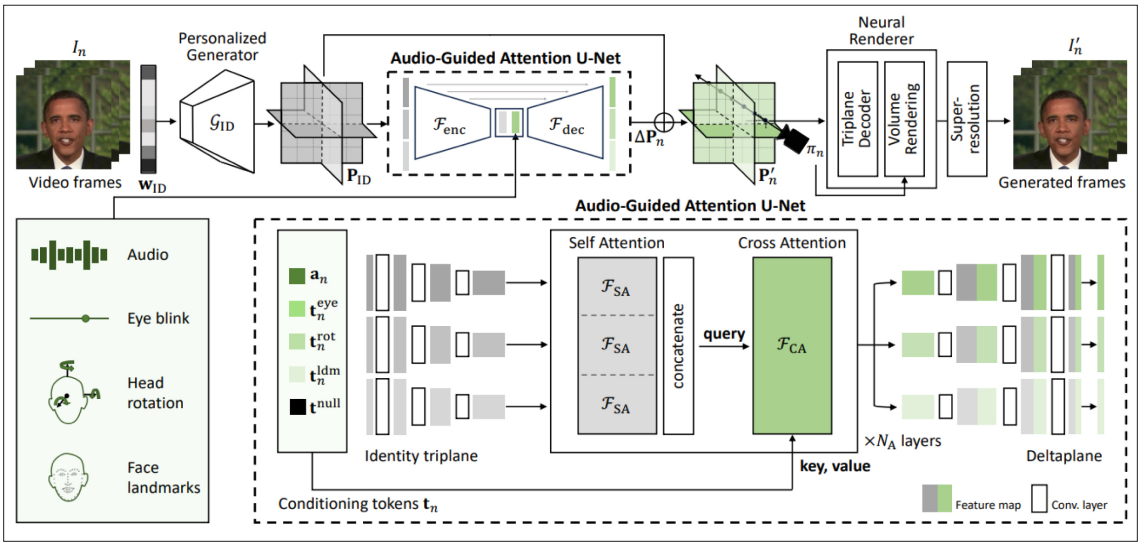
$$I_n^{\mathrm{novel}} = \mathcal{R}(\mathbf{P}', \pi_n; \theta^*_{\mathcal{R}}) = \mathcal{R}(\mathbf{P}_{\mathrm{ID}} + \Delta\mathbf{P}_n^{\mathrm{novel}}, \pi_n; \theta^*_{\mathcal{R}}). \tag{3}$$

다음으로, 우리는 오디오 기반의 deltaplane 예측 방법과, 사용되는 손실 함수에 대해 설명한다. 제안된 방법의 개요는 〈그림 2〉에 도식화되어 있다.

2. 손실 함수

모델을 훈련하기 위해 우리는 주로 L1손실과 LPIPS 손실[19]을 사용하여 입력 프레임 I를 재구성한다. 또한, 얼굴 영역의 시멘틱 분할을 BiSeNet[20]을 통해 얻어, 입술 세그먼트에 대한 reconstruction 손실을 추가한다. 추가적으로, ID 유사성 손실과 SyncNet 손실 함수[21,22]를 사용하여 생성 결과를 최적화한다. 요약하면, 전체 손실 함수 L은 다음과 같이 정의된다:

또한, 우리는 super-resolution 모듈을 업데이트하기



<그림 2>

$$\mathcal{L} = \mathcal{L}_{\text{rec}}(I, I') + \lambda_{\text{lip}}\mathcal{L}_{\text{lip}}(S_{\text{lip}}(I), S_{\text{lip}}(I')) + \lambda_{\text{id}}\mathcal{L}_{\text{id}}(I, I') + \lambda_{\text{sync}}\mathcal{L}_{\text{sync}}(I, I'). \quad (4)$$

위해 수 에포크를 추가로 미세 조정한다. 이 단계는 더욱 선명한 이미지를 생성하기 위함이며, 이때에는 재구성 손실만을 적용한다.

Ⅲ. 개인화된 오디오 기반 NeRF 기술의 실험 결과

오디오 기반 발화 영상 합성을 위해, 오디오 트랙과 함께 몇 분간의 말하는 초상화 비디오가 필요하다. 우리는 기존 논문에서 활용한 데이터셋[14]을 사용한다. 이 데이터셋은 평균 6,000프레임(25fps)의 인물 중심 비디오로 구성된다. 이전의 NeRF 기반 방법론[14,15,16,17]을 따라 비디오를 학습용과 테스트용 세트로 나눈다. 또한, Wav2Vec 모델[23]을 사전 학습된 상태로 사용하여 각 오디오에서 feature를 추출한다.

우리의 방법을 Wav2Lip[22]과 PC-AVS[24]와 같은 2D 기반의 발화 영상 생성 모델과 비교한다. 또한, AD-NeRF[14], RAD-NeRF[15], ER-NeRF[17] 등의 여러 NeRF 기반 모델과도 비교한다. 공정한 비교를 위해, 생성된 결과를 동일한 크기 영역으로 얼굴 영역을 잘라내고 크기를 조정한다. 또한, NeRF 기반 방법들이 비디오에서 추출된

<표 1>

Head angle (yaw, pitch)	(−30°, −20°)			(−15°, −10°)			(+15°, +10°)			(+30°, +20°)		
	Sync↑	FID↓	IDSIM↑	Sync↑	FID↓	IDSIM↑	Sync↑	FID↓	IDSIM↑	Sync↑	FID↓	IDSIM↑
AD-NeRF [23]	2.236	212.845	0.068	3.474	175.978	0.280	3.821	152.018	0.481	2.523	193.343	0.034
RAD-NeRF [48]	4.938	167.834	0.186	5.543	123.924	0.378	6.831	94.674	0.607	5.447	185.718	0.283
ER-NeRF [30]	4.774	198.291	0.226	7.335	87.594	0.575	6.652	80.562	0.503	2.702	141.625	0.022
Talk3D (Ours)	7.201	81.113	0.611	7.932	37.774	0.766	8.144	39.971	0.797	7.766	68.680	0.643

GT 배경을 그대로 사용하기 때문에, 생성된 결과의 얼굴 영역에서만 재구성 지표를 측정한다.

1. 정량적 평가

정량적 평가에서는 재구성 품질과 입술 동기화 정확도를 세 가지 다른 설정: 1) novel-view synthesis 실험, 2) self-driven 실험, 3) cross-driven 실험에서 비교한다. 첫 번째 novel-view synthesis은 다양한 새로운 카메라 뷰포인트에서 렌더링하여 다양한 카메라 각도에서의 렌더링 성능을 평가한다. 구체적으로, 기준 뷰포인트에서 일정 각도(좌우 30도, 위아래 20도)를 변경하며 수행한다. Self-driven은 렌더링 카메라 뷰포인트만 실제 비디오에서 추출된다는 점을 제외하고는 novel-view synthesis 설정과 동일하다. 마지막으로, cross-driven 설정에서는 완전히 무관한 오디오 데이터에 의해 생성된 lip-sync 정확도를 SynObama[25]의 데모에서 추출한 두 개의 오디오 클립을 사용하여 측정한다.

이미지 재구성 품질을 평가하기 위해, PSNR, SSIM, LPIPS를 사용한다. 또한, FID, LMD, AUE, lip-sync accuracy(Sync) 지표를 사용하여 이미지 퀄리티와 입술 동기화 정확도를 평가한다. 또한, 얼굴 identity의 유사성을 측정하기 위해 identity 유사성 지표(IDSIM)를 사용한다.

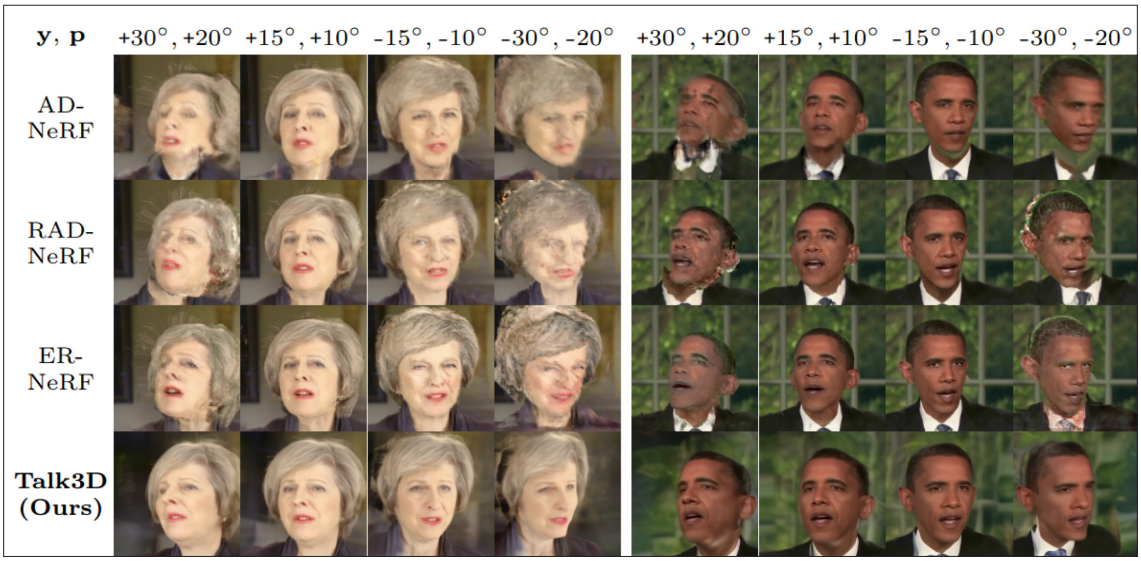
Novel-view synthesis 실험의 결과는 〈표 1〉에 나와 있다. 여기서는 카메라 뷰포인트를 직접적으로 제어할 수 있는 이전 작업들과 우리 방법을 비교한다. 대부분의 방법들이 정면 뷰 렌더링에서는 우리 모델과 비슷한 성능을 보이지만, 카메라 뷰포인트가 극단적인 각도로 회전될 때 점수가 크게 떨어진다. 반면, 우리 방법은 모든 지표에서 일관되게 높은 점수를 보여주며, 다양한 카메라 뷰포인트에서도 생성 품질과 입술 동기화 정확도를 유지하는 데 효

<표 2>

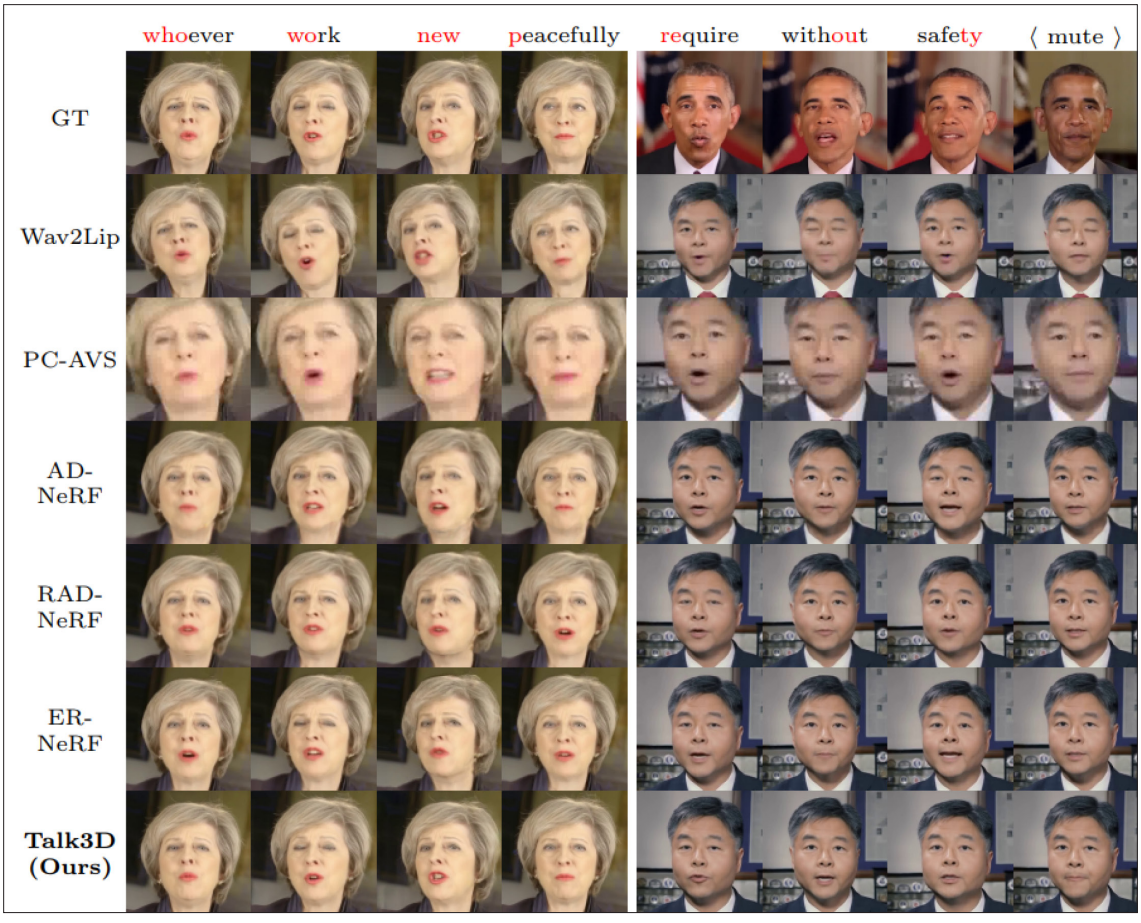
Methods	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	LMD ↓	AUE ↓	Sync ↑	IDSIM ↑
Ground Truth	N/A	N/A	0	0	0	0	9.077	1
Wav2Lip [39]	28.678	0.862	0.053	33.074	4.658	3.040	10.096	0.893
PC-AVS [69]	20.729	0.638	0.112	42.646	3.419	2.497	8.945	0.520
AD-NeRF [23]	27.611	0.877	0.049	20.243	5.692	2.331	5.692	0.904
RAD-NeRF [48]	28.797	0.886	0.038	14.218	3.467	2.163	6.316	0.921
ER-NeRF [30]	29.284	0.891	0.032	11.860	3.417	2.025	6.724	0.940
Talk3D(Ours)	30.185	0.895	0.027	8.626	2.932	1.920	7.383	0.944

<표 3>

	Testset A			Testset B		
Methods	Sync↑	LMD↓	AUE↓	Sync↑	LMD↓	AUE↓
Ground Truth	7.850	0	0	6.976	0	0
Wav2Lip [39]	8.272	7.039	4.154	7.907	5.561	3.967
PC-AVS [69]	8.408	7.754	6.278	7.533	6.560	4.518
AD-NeRF [23]	5.670	7.378	4.736	5.076	5.542	3.711
RAD-NeRF [48]	6.532	5.848	4.717	5.472	5.599	3.666
ER-NeRF [30]	6.507	6.181	4.489	5.160	5.374	3.519
Talk3D (Ours)	6.827	5.352	4.693	5.780	4.814	3.132


y, p
+30°, +20°
+15°, +10°
-15°, -10°
-30°, -20°
+30°, +20°
+15°, +10°
-15°, -10°
-30°, -20°
AD-NeRF
RAD-NeRF
ER-NeRF
Talk3D (Ours)

<그림 3>


whoever
work
new
peacefully
require
without
safety
⟨ mute ⟩
GT
Wav2Lip
PC-AVS
AD-NeRF
RAD-NeRF
ER-NeRF
Talk3D (Ours)

<그림 4>

과적임을 보여준다.

Self-driven 평가 결과는 〈표 2〉에 제시되어 있다. 우리 방법은 대부분의 이미지 품질 지표에서 최고의 품질을 달성하며, NeRF 기반 방법 중에서도 최고의 입술 동기화를 보여준다. 단일 샷 2D 기반 방법인 Wav2Lip과 PC-AVS는 뛰어난 Sync 점수를 보이지만, 이미지 충실도에서 낮은 점수를 보이는 것으로 이미지의 정확한 재구성에 적합하지 않음을 보여준다. 우리 방법은 PSNR, SSIM, LPIPS, 그리고 ID-SIM 점수에서 우수한 성능을 보여 고품질의 이미지 재구성에 효과적이고, 또한 얼굴 identity를 유지할 수 있음을 입증한다. 또한, 높은 FID 점수는 3D GAN의 prior를 활용하는 것이 재구성에 특화된 기존 방법들에 비해 FID 측면에서 더 유리하다는 것을 나타낸다. LMD와 AUE에서의 뛰어난 점수는 우리 방법이 얼굴 역학을 더 정확하게 나타내는 발화 영상을 생성한다는 것을 보여준다.

Cross-driven 평가 결과는 〈표 3〉에 나와 있으며, 일반 오디오 입력으로부터 해당 입술 움직임을 성공적으로 합성하는 성능을 보여준다. 우리 모델은 NeRF 기반 방법들 중 대부분의 비교에서 일관되게 최고 점수를 보여준다. 추가적으로 사용한 Sync 손실 함수는 모델이 학습 중 보지 못한 오디오에서도 정확한 입술 모양을 생성하도록 유도한다. 이는 Sync 손실을 사용할 수 없는 이전 모델들과 확연히 구별되는 점이다.

2. 정성적 평가

이 섹션에서는 각 평가 설정에서 생성된 결과를 보여준다. 먼저 다양한 뷰포인트에서 렌더링된 얼굴 이미지를 시각화한다. 〈그림 3〉에서 볼 수 있듯이, 이전의 NeRF 기반 방법들은 정면 뷰포인트에서 멀리 떨어진 카메라 각도에서 생성된 경우 성능 저하를 겪으며, 불규칙한 얼굴 색상과 아티팩트를 자주 보여준다. 특히, RAD-NeRF와 ER-NeRF는 pseudo-3D deformable module이라는 그들의 몸통 모델링 때문에 부적절한 몸통 생성 결과를 보여준다. AD-NeRF는 머리와 몸통 볼륨을 독립적으로 학습하므로, 측면 각도에서 보았을 때 생성된 머리가 분리된다.

또한, self-driven 및 cross-driven 실험에서 샘플링된 결과를 〈그림 4〉에서 보여준다. 각 실험에서 네 개의 주요 프레임을 선택하여 입술 동기화 정확도와 재구성 품질을 비교하였다. Wav2Lip과 PC-AVS와 같은 2D 기반 방법들은 높은 입술 동기화 정확도를 보여주지만, 생성된 결과는 주어진 장면을 정확히 재구성할 수 없다. 또한 NeRF 기반 방법들은 분리된 렌더링 파이프라인은 목 부분에서 부자연스러운 머리 움직임을 보이며, 특히 머리카락 부분에서 흐릿한 텍스처를 보여준다. 반면, 우리 Talk3D는 통합된 생성 프로세스와 3D GAN의 prior 덕분에 더욱 정확한 결과를 보여주며, 목 부분에서의 부자연스러운 생성 또한 발생하지 않는다.

Ⅳ. 결 론

본 논문에서는 3D 생성형 모델 및 고품질의 3D 발화 영상 생성을 위한 새로운 프레임워크인 Talk3D를 소개했다. 우리의 프레임워크는 개인화된 3D GAN을 오디오 기반 발화 영상 생성과 통합하여 현실적인 얼굴의 geometry를 유지하는 동시에 카메라 방향 제어를 통해 실제적인 3D 발화 아바타를 생성할 수 있다. 또한, 우리가 제안한 오디오 가이드 attention U-Net 아키텍처는 배경, 몸통 및 눈 움직임과 같은 이미지 프레임 내의 국지적인 변형의 disentanglement를 향상시킨다. 우리는 광범위한 실험을 통해 제안된 모델이 입력 오디오에 해당하는 정확한 입술 움직임을 생성할 뿐만 아니라 새로운 관점에서 렌더링을 가능하게 하여 이전의 모델들과 비교하였을 때 더욱 우수한 발화 영상을 만들어 낸다는 것을 입증한다. 우리의 모델이 디지털 미디어 매체와 가상 현실 연구와 영화 제작 및 화상 회의에서 큰 도움이 될 것을 희망한다.

참 고 문 헌

[1] Chan, Eric R., et al. “Efficient geometry-aware 3d generative adversarial networks.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[2] An, Sizhe, et al. “Panohead: Geometry-aware 3d full-head synthesis in 360deg.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.
[3] Sun, Jingxiang, et al. “Next3d: Generative neural texture rasterization for 3d-aware head avatars.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.
[4] Ko, Jaehoon, et al. “3d gan inversion with pose optimization.” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[5] Bhattarai, Ananta R., Matthias Nießner, and Artem Sevastopolsky. “Triplanenet: An encoder for eg3d inversion.” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.
[6] Trevithick, Alex, et al. “Real-time radiance fields for single-image portrait view synthesis.” (2023).
[7] Bai, Yunpeng, et al. “High-fidelity facial avatar reconstruction from monocular video with generative priors.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[8] Frühstück, Anna, et al. “VIVE3D: Viewpoint-independent video editing using 3D-aware GANs.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[9] Mildenhall, Ben, et al. “Nerf: Representing scenes as neural radiance fields for view synthesis.” Communications of the ACM 65.1 (2021): 99-106.
[10] Müller, Thomas, et al. “Instant neural graphics primitives with a multiresolution hash encoding.” ACM transactions on graphics (TOG) 41.4 (2022): 1-15.
[11] Fridovich-Keil, Sara, et al. “Plenoxels: Radiance fields without neural networks.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[12] Wang, Tengfei, et al. “Rodin: A generative model for sculpting 3d digital avatars using diffusion.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.
[13] Chen, Xingyu, Yu Deng, and Baoyuan Wang. “Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation.” 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, 2023.
[14] Guo, Yudong, et al. “Ad-nerf: Audio driven neural radiance fields for talking head synthesis.” Proceedings of the IEEE/CVF international conference on computer vision. 2021.
[15] Tang, Jiaxiang, et al. “Real-time neural radiance talking portrait synthesis via audio-spatial decomposition.” arXiv preprint arXiv:2211.12368 (2022).
[16] Liu, Xian, et al. “Semantic-aware implicit neural audio-driven video portrait generation.” European conference on computer vision. Cham: Springer Nature Switzerland, 2022.
[17] Li, Jiahe, et al. “Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
[18] Ekman, Paul and Wallace V. Friesen. “Facial Action Coding System: Manual.” (1978).
[19] Zhang, Richard, et al. “The unreasonable effectiveness of deep features as a perceptual metric.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[20] Yu, Changqian, et al. “Bisenet: Bilateral segmentation network for real-time semantic segmentation.” Proceedings of the European conference on computer vision (ECCV). 2018.
[21] Chung, Joon Son and Andrew Zisserman. “Out of Time: Automated Lip Sync in the Wild.” ACCV Workshops (2016).
[22] Prajwal, K. R., et al. “A lip sync expert is all you need for speech to lip generation in the wild.” Proceedings of the 28th ACM international conference on multimedia. 2020.
[23] Baevski, Alexei, et al. “wav2vec 2.0: A framework for self-supervised learning of speech representations.” Advances in neural information processing systems 33 (2020): 12449-12460.
[24] Zhou, Hang, et al. “Pose-controllable talking face generation by implicitly modularized audio-visual representation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[25] Suwajanakorn, Supasorn et al. “Synthesizing Obama.” ACM Transactions on Graphics (TOG) 36 (2017): 1 - 13.

저 자 소 개

고 재 훈

- 2015년 3월 ~ 2021년 8월 : 고려대학교 이과대 지구환경과학과 (학부 졸업)
- 2022년 3월 ~ 2024년 2월 : 고려대학교 정보대학 컴퓨터보안전공 (석사과정 졸업)

김 승 룡

- 2008년 3월 ~ 2012년 2월 : 연세대학교 전기전자공학과 학사
- 2012년 3월 ~ 2018년 2월 : 연세대학교 전기전자공학과 박사
- 2018년 3월 ~ 2019년 2월 : 연세대학교 전기전자공학과 박사후 연구원
- 2019년 3월 ~ 2020년 2월 : 스위스 EPFL 박사후 연구원
- 2020년 3월 ~ 2024월 2월 : 고려대학교 컴퓨터학과 조교수
- 2024년 3월 ~ 현재 : 고려대학교 컴퓨터학과 부교수