Attention Mesh: High-fidelity Face Mesh Prediction in Real-time

해당 논문은 어텐션 메쉬(Attention Mesh)를 제안한다. 이는 의미 있는 영역을 위해 어텐션을 사용하는 3D 얼굴 메쉬를 예측하는 경량 아키텍처로, 실시간 장치 내 추론을 위해 설계된 신경망이다. Google Pixel 2에서 초당 50 프레임 이상으로 실행된다. 제안된 방법론은 AR 메이크업, 시선 추적 및 AR 퍼펫팅(puppeteering)과 같은 응용 프로그램을 가능하게 하며, 눈과 입 주변의 높은 정확도의 랜드마크에 의존하는 기능들을 제공한다. 주요 기여는 다단계 계층 접근 방식과 동등한 정확도를 달성하면서도 30% 빠른 통합된 네트워크 아키텍처이다.

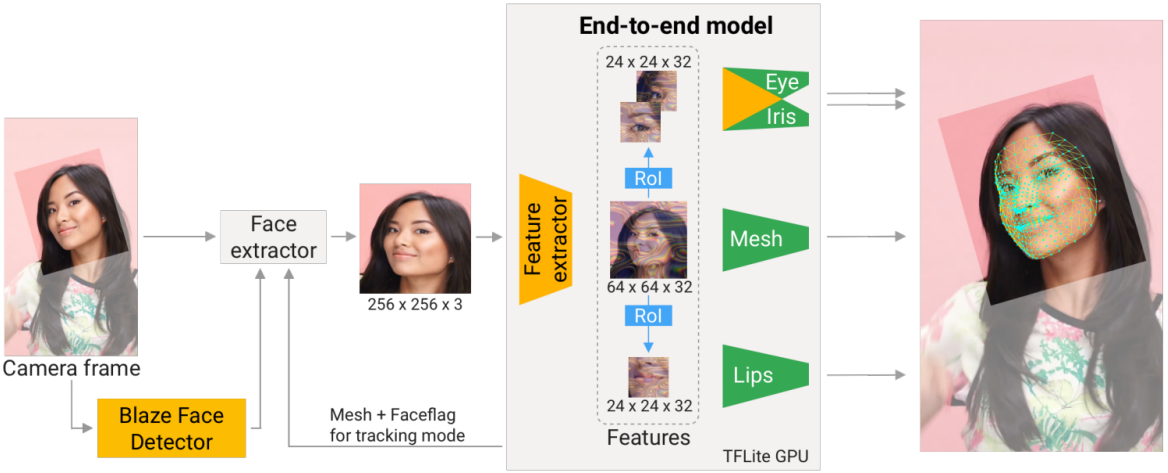
이 연구에서는 이미지상의 인간 얼굴에 자세한 3D 메쉬 템플릿을 등록하는 문제에 대해 다룬다. 등록된 메쉬는 립스틱 가상 시착이나 가상 아바타의 퍼펫팅과 같이 입술과 눈 윤곽의 정확도가 현실감에 중요한 가상 시뮬레이션에 사용될 수 있다.

인간 얼굴의 매개변수 모델을 사용하는 방법과 달리, 우리는 3D 얼굴 메쉬 정점의 위치를 직접 예측한다. 해당 논문은 이 분야의 이전 연구를 기반으로 하여 얼굴 감지기 다음에 랜드마크 회귀 네트워크가 따르는 두 단계 아키텍처를 사용한다. 그러나 전체 얼굴에 대해 단일 회귀 네트워크를 사용하면 감각적으로 더 중요한 지역(입술, 눈)에서 품질이 저하된다.

<그림 1: 어텐션 메쉬 서브모델로 예측된 돌출 굴곡>

이 문제를 완화하는 한 가지 방법은 단계적인 접근 방식이다. 초기 메쉬 예측을 사용하여 이러한 지역 주위에 촘촘한 영역을 생성하고 이를 전문화된 네트워크에 전달하여 높은 품질의 랜드마크를 생성한다. 이는 정확도 문제를 직접 해결하지만, 원본 이미지를 입력으로 사용하는 상대적으로 큰 독립된 모델 및 GPU와 CPU 간의 매우 비용이 많이 드는 추가 동기화 단계를 도입한다. 해당 논문에서는 이러한 단계적인 접근 방식과 동등한 품질을 달성하는 단일 모델이 가능하다는 것을 보여준다. 이는 공간 변형기를 사용하여 특정 영역의 특징 맵을 변환하는 영역별 헤드를 사용하면서 추론 중에 최대 30% 빠를 수 있다. 해당 논문은 제안된 아키텍처를 어텐션 메쉬(Attention Mesh)라고 부른다. 추가적인 이점은 여러 이질적인 네트워크 대신 내부적으로 일관된 훈련 및 배포가 가능하다.

우리는 다양한 얼굴 감지기에서 제공된 초기화에 강건한 네트워크를 구축하는 논문에서 설명된 것과 유사한 아키텍처를 사용하고 있다. 두 논문의 목표가 다르더라도, 공간 변형기를 사용하여 현저한 개선을 이끌어내는 동안 주요 얼굴 영역에 해당하는 헤드와 결합하는 것이 단일 대형 네트워크보다 효과적이라는 점은 흥미로운 사실이다. 눈, 홍채 그리고 입술에 해당하는 랜드마크를 생성하는 해당 논문의 구현 세부 사항과 품질 및 추론 성능 벤치마크를 제공한다.



<그림 2: 추론 파이프라인과 모델 아키텍처 개요>

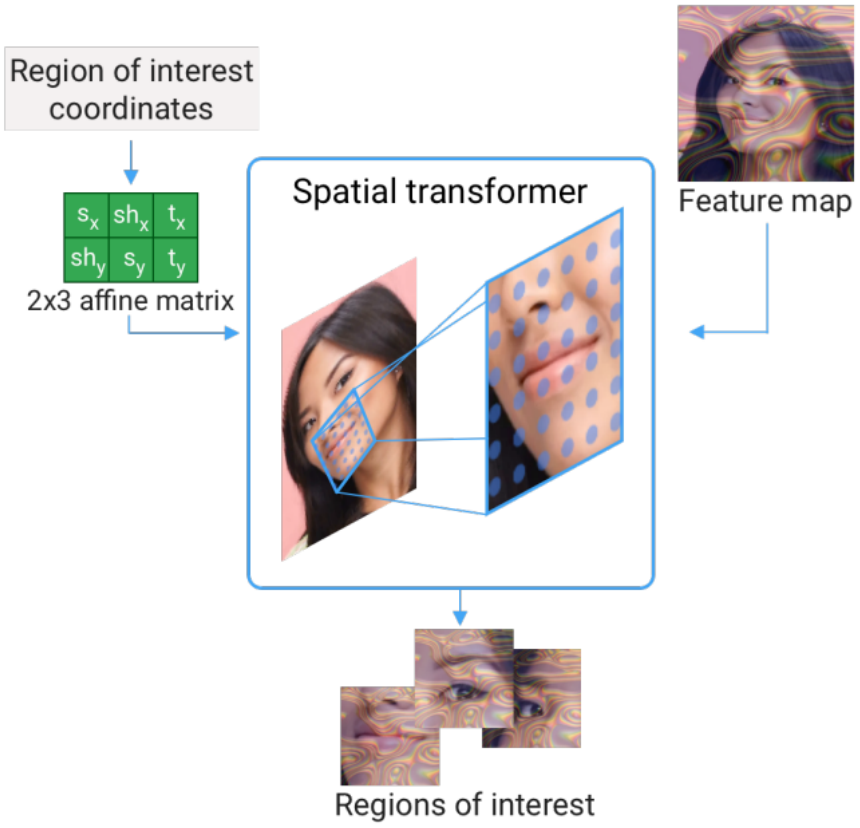
제안된 모델은 256 × 256 이미지를 입력으로 받는다. 이 이미지는 얼굴 감지기 또는 이전 프레임에서의 추적을 통해 제공된다. 64 × 64 특징 맵을 추출한 후 모델은 여러 하위 모델로 분할된다. 하나의 하위 모델은 3D 얼굴 메쉬의 478개의 랜드마크를 모두 예측하고 각 관심 영역에 대한 자르기 경계선을 정의한다. 나머지 하위 모델은 어텐션 메커니즘을 통해 얻은 해당 24×24 특징 맵에서 지역 랜드마크를 예측한다.

해당 논문은 주요 윤곽이 있는 세 가지 얼굴 영역에 중점을 둔다(입술 및 두 눈). 각 눈 하위 모델은 공간 해상도 6 × 6에 도달한 후 분리된 출력으로 홍채를 예측한다. 이렇게 하면 눈 특징을 재사용하면서 동적인 홍채를 정적인 눈 랜드마크와 독립적으로 유지할 수 있다.
개별 하위 모델을 사용하면 각 영역에 할당된 네트워크 용량을 제어하고 필요한 곳에서 품질을 향상시킬 수 있다. 예측의 정확도를 더 향상시키기 위해 눈과 입술이 수평에 정렬되어 있고 균일한 크기를 가지도록 일련의 정규화를 적용한다.

시각적 특징 추출을 위해 여러 어텐션 메커니즘(소프트 및 하드)이 개발되었다. 이러한 어텐션 메커니즘은 특징 공간에서 2D 포인트 그리드를 샘플링하고 샘플링된 포인트 하단의 특징을 미분 가능한 방식으로 추출한다(2D 가우시안 커널 또는 어파인 변환 및 미분 가능한 보간 사용). 이를 통해 엔드 투 엔드로 아키텍처를 훈련하고 어텐션 메커니즘에서 사용되는 특징을 풍부하게 만들 수 있다. 구체적으로는 64 × 64 특징 맵에서 24 × 24 지역 특징을 추출하기 위해 공간 변형 모듈을 사용한다. 공간 변형 모듈은 어파인 변환 행렬 θ (식 1)에 의해 제어되며 포인트 그리드를 확대, 회전, 이동 및 기울일 수 있다.

$$\theta = \begin{bmatrix} x_x & sh_x & t_x \\ sh_y & s_y & t_y \end{bmatrix} \tag{1}$$

이 어파인 변환은 매트릭스 매개변수의 지도 예측을 통해 구성되거나 얼굴 메쉬 하위 모델의 출력에서 계산될 수 있다.



<그림 3: 어텐션 메커니즘을 사용한 공간 변형기>

통합된 접근 방식을 평가하기 위해 기본 메쉬, 눈 및 입술에 대한 독립적으로 훈련된 지역별 모델의 연속된 단계적인 접근 방식 모델과 비교한다.

표 1은 전형적인 최신 모바일 장치에서 얼굴 및 지역 모델의 단계적인 접근 방식 얼굴 분할보다 어텐션 메쉬가 25% 이상 빠르게 실행됨을 보여준다. 성능은 TFLite GPU 추론 엔진을 사용하여 측정되었다. GPU에서 전체 어텐션 메쉬 추론이 한 번에 수행되기 때문에 비용이 많이 드는 CPU-GPU 동기화의 감소로 추가적인 5% 속도 향상이 이루어진다.

Model	Inference Time (ms)
Mesh	8.82
Lips	4.18
Eye & iris	4.70
Cascade (sum of above)	22.4
Attention Mesh	**16.6**

<표 1: Pixel 2XL에서의 성능>

두 모델의 정량적인 비교가 표 2에 제시되어 있다. 대표적인 메트릭으로는 특정 지점의 예측된 위치와 실제 위치 간의 평균 거리를 사용하며, 이를 3D 두 눈 사이의 거리 또는 입술과 눈의 경우 모서리 간의 거리로 정규화하여 크기 불변성을 보장한다. 어텐션 메쉬 모델은 눈 영역에서 모델의 단계적인 접근 방식 모델을 능가하며, 입술 영역에서는 비슷한 성능을 보여준다.

Model	All	Lips	Eyes
Mesh	2.99	3.28	6.6
Cascade	2.99	2.70	6.28
Attention mesh	**3.11**	**2.89**	**6.04**

<표 2: 2D에서 평균 정규화 오차>