한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 32, No. 3, P. 1~11

기하학적 변형과 딥러닝 복원을 결합한 게임 스프라이트 생성 방법

오영택

호서대학교 게임소프트웨어학과

ytoh@hoseo.edu

Game Sprite Generation via Geometric Deformation and Deep Restoration

Young-Taek Oh

Department of Game Software, Hoseo University

요 약

본 논문은 게임 스프라이트 애니메이션 생성에서 발생하는 프레임 간 외형 일관성 부족과 포즈 제어의 어려움을 해결하기 위해, 기하학적 변형과 딥러닝 복원을 결합한 방법을 제안한다. 먼저 입력 스프라이트로부터 변형에 적합한 2차원 메쉬를 구성하고, 사용자 관절 조작을 메쉬 정점 제약으로 직접 반영하는 ARAP 기반 변형을 통해 목표 포즈에 대응하는 변형 이미지를 생성한다. 이후 비대칭 U-Net(AsymUNet) 기반 복원 네트워크가 변형 과정에서 발생한 블러, 이미지 결손, 국소 왜곡을 보정하여 원본 외형과 일관성 있는 스프라이트를 복원한다. 제안 방법은 관절 기반의 직관적인 포즈 제어를 통해 안정적인 외형 일관성을 유지하며, 저해상도 도트 스프라이트에도 적합한 파이프라인을 제공한다.

Abstract

This paper proposes a method for game sprite animation that combines geometric deformation with deep restoration to address appearance inconsistency and limited pose controllability in sprite creation. First, a 2D mesh suitable for deformation is constructed from an input sprite, and user-specified joint positions are directly applied as mesh-vertex constraints via ARAP-based deformation to produce a warped frame matching the target pose. An asymmetric U-Net (AsymUNet) then refines this warped output by correcting blur, image defects, and local distortions introduced during warping, while preserving the original character appearance. The proposed pipeline offers intuitive joint-based pose control and stable appearance consistency, and is well suited to low-resolution pixel-art sprites.

키워드: 스프라이트 생성, 기하학적 변형, ARAP, 비대칭 U-Net, 픽셀 아트 복원
Keywords: Sprite generation, Geometric deformation, ARAP, Asymmetric U-Net, Pixel-art restoration

1 서론

게임 스프라이트 애니메이션은 캐릭터 동작을 표현하는 연속 프레임으로 구성되며, 프레임 간 외형 일관성을 유지하면서 다양한 포즈를 구현하는 것이 제작 과정의 핵심 과제이다. 그러나 이러한 스프라이트를 직접 제작하는 과정은 많은 시간과 노력이 요구된다.

최근 생성형 인공지능 기술의 비약적인 발전으로, Stable Diffusion 기반 도구들을 활용하면 간단한 프롬프트만으로 원하는 게임 스프라이트를 생성할 수 있게 되었다. 특히 단일 레퍼런스 이미지로부터 연속적인 포즈 프레임을 생성하는 연구들이 등장하면서, 스프라이트 애니메이션 자동화에 대한 기대도 높아졌다.

그러나 걷기와 같은 게임용 스프라이트 애니메이션 제작에는 여전히 적용하기 어렵다. 이는 생성형 인공지능 기반 영상 생성 기법이 프레임 간 외형 일관성을 충분히 유지하지 못하거나, 사용자가 의도한 포즈를 정확하게 반영하지 못하기 때문이다.

이러한 문제를 해결하기 위해 ReferenceNet을 활용하는 Animate Anyone [1] 및 Sprite Sheet Diffusion [2]과 같은 연구가 진행되었으나, 캐릭터 외형의 일관성 문제를 완전히 해결하지는 못

*correspondin author: Young-Taek Oh / Department of Game Software, Hoseo University (ytoh@hoseo.edu)

Received : 2026.05.22./ Review completed : 1st 2026.06.07. 2nd 2026.06.25./ Accepted : 2026.06.29.
DOI : 10.15701/kcgs.2026.32.3.1
ISSN : 1975-7883(Print)/2383-529X(Online)

용하여 사람이나 인간형 캐릭터를 애니메이션화하는 방법을 제안하였다. 사용자가 입력 이미지에서 관절 위치를 지정하면, 이에 대응하는 카메라와 3차원 자세를 복원하고, 계층 구조를 갖는 삼각형 메쉬 템플릿을 캐릭터에 맞게 정렬하였다. 이후 ARAP 변형 기법을 기반으로, 원근 왜곡 효과를 반영한 ASAP(As-Similar-As-Possible) 변형을 적용하여 3차원 모션에 따른 자연스러운 2차원 캐릭터 애니메이션을 생성하였다.

3 제안 방법

3.1 전체 시스템 구조

시스템은 크게 두개의 부분으로 구성된다. 먼저 ARAP기반의 스프라이트 변형 모듈은 입력 이미지를 기하학적 변형 방법을 활용하여 사용자가 의도한 포즈를 만들어 낸다. 다음으로 복원 네트워크 기반 품질 보정 모듈은 변형된 이미지와 변형에 사용된 레퍼런스 이미지를 입력으로 하여 최대한 원본 이미지와 유사한 스프라이트를 생성해 낸다.

3.2 ARAP 기반 스프라이트 변형

ARAP 기반 스프라이트 변형 모듈은 입력 전처리와 메쉬 생성, 스켈레톤 리깅, 포즈 편집 및 변형 단계로 구성된다. 여기서는 원본 이미지로부터 생성된 2차원 메쉬에 구조적 제어를 가하여 복원 네트워크에 들어갈 입력 이미지를 만들어 내는 것을 목표로 한다. 이를 위하여 관절 제어에 따라 국소 강체성을 최대한 유지하는 변형 메쉬를 계산하고, 이를 바탕으로 타겟 포즈와 최대한 유사한 이미지를 만든다.

Figure 2은 메쉬 생성(mesh tool), 스켈레톤 리깅(skeleton tool), 포즈 편집(pose tool)으로 이어지는 편집 도구의 3단계 작업 흐름을 보여준다. (a) mesh tool에서는 전경 마스크와 절단 경로를 반영해 변형에 적합한 삼각형 메쉬를 구성한다. (b) skeleton tool에서는 메쉬 위에 관절과 뼈대 연결을 배치하여 변형 제어 구조를 정의한다. (c) pose tool에서는 관절을 목표 위치로 조작해 ARAP 제약을 적용하고, 그 결과로 생성된 변형 프레임을 확인한다.

3.2.1 기하학적 변형 이미지 생성

3.3.1절에서 사용하는 레퍼런스 이미지는 32×32 RGBA 프레임이므로, 이를 그대로 사용하면 ARAP 기반 기하학적 변형 과정에서 국소 제어가 어렵다. 따라서 입력 이미지를 최근접 이웃 보간(Nearest-Neighbor Interpolation)으로 128×128까지 업스케일한 뒤 변형을 수행하며, ARAP 모듈의 출력 역시 128×128 해상도로 생성한다.

3.2.2 메쉬 생성

레퍼런스 이미지로부터 전경 영역을 추출한 뒤, 이를 바탕으로 변형 가능한 2차원 삼각형 메쉬를 구성한다. 이때 중요한 점은 스프라이트에서 하나로 붙어 보이는 구조와, 변형에 적합한 위상 구조가 반드시 일치하지 않는다는 것이다. 예를 들어 팔과 몸통, 다리와 몸통이 하나의 실루엣으로 붙어 있는 경우, 이를 그대로 삼각분할을 하게 되면 이후 관절 변형 시 서로 다른 신체 부위가 동일한 삼각형 패치를 공유하게 되어 부자연스러운 변형이 발생한다.

이를 해결하기 위해 메쉬 생성 단계에서는 사용자가 지정한 절단 경로를 반영하여 마스크의 위상을 먼저 보정한다. 절단 경로는 LiveWire [11] 계열의 에지 추적 방식을 따라 영상 경계에 정렬되도록 계산되며, 적용 후에는 붙어 있던 부위를 분리하거나 내부 제거 구간을 형성한다. 이때 절단 입력을 삼각분할 이후의 후처리로 다루지 않고, 삼각분할 단계에서 제외되는 띠 형태의 다각형 구멍(polygonal hole)으로 변환하였다. 이는 절단 경계가 메쉬 생성 과정에 직접 반영되도록 하여, 단순한 삼각형 삭제보다 더 안정적인 위상 분리를 가능하게 한다. 이후 수정된 경계에 대해 제약 삼각분할(constrained triangulation)을 수행하여 최종 메쉬를 생성한다. 이 과정은 단순한 메쉬 분할이 아니라, 후속 포즈 변형에 적합한 변형 친화적 위상 구조(deformation-friendly topology)를 구성하는 단계로 볼 수 있다.

3.2.3 스켈레톤 리깅

생성된 메쉬 위에는 인체형 2차원 스켈레톤 $\mathcal{S} = (J, B)$를 정의한다. 여기서 $J = \{j_m\}$는 관절 집합이고, B는 관절 간 연결 관계를 나타내는 뼈대 그래프(bone graph)이다. 관절은 골반, 목, 머리, 양측 어깨, 팔꿈치, 손목, 무릎, 발목으로 구성되며, 이는 스프라이트 포즈 편집에 필요한 최소한의 구조적 자유도를 제공한다.

관절은 처음부터 메쉬 정점에 고정하지 않고, 연속적인 2차원 공간에서 위치를 조정한 뒤 변형 계산 단계에서 메쉬 정점으로 삽입한다. 이때 각 관절은 자신이 포함된 삼각형을 분할하여 새 정점으로 추가된다. 따라서 관절 조작은 메쉬 정점에 대한 직접 제약으로 해석되며, 그 효과가 ARAP 변형에 즉시 반영된다.

3.2.4 포즈 편집 및 변형

포즈 편집 단계에서는 스켈레톤 제어를 통해 제어점 집합 Q를 목표 위치 Q'로 이동시키고, 이를 만족하는 변형 메쉬를 계산한다. 손목과 발목은 말단 관절로 간주하여 2절 역기구학(two-bone inverse kinematics)을 적용하고, 나머지 관절은 순기구학(forward kinematics) 방식으로 제어한다. 예를 들어 루트 관절 r, 중간 관절 m, 말단 관절 e로 이루어진 체인에 대해 목표 위치 t가 주어지면, 새로운 중간 관절 m'은 뼈 길이 보존 조건 $\|m' - r\| = L_1$, $\|t - m'\| = L_2$ 을 만족하도록 계산된다. 여기서 L_1, L_2는 기준 자세(rest pose)에서의 뼈 길이이다.

이후 변형 메쉬는 ARAP 에너지 최소화를 통해 계산된다. 제어점 집합이 정해지면 ARAP 최적화에 필요한 선형 시스템을 미리 분해해 두고, 이후에는 목표 관절 위치 변화만 반영하여 빠르게 해를 반복 계산한다. 즉, 각 삼각형의 국소 강체성을 최대한 유

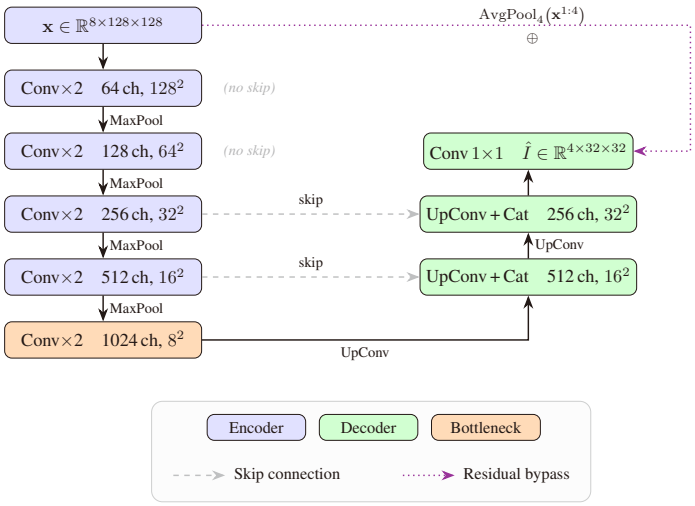


Figure 3: Asymmetric UNet (AsymUNet) architecture

참조 이미지를 입력에 포함하는 설계는 예비 실험으로부터 도출되었다. 변형 이미지만을 4채널 입력으로 사용하면, 모델이 ARAP 변형이 발생하지 않은 영역에도 불필요한 변환을 가하는 경향이 관찰되었다. 반면, 변형 전 원본인 참조 이미지를 추가 입력으로 제공하면 변형된 영역과 그렇지 않은 영역의 구분이 가능해지며, 변형 영역에 복원을 집중함으로써 픽셀 아트의 선명도와 지각 품질이 향상됨을 확인하였다. 이에 따라 변형 이미지와 참조 이미지를 채널 방향으로 이어붙인 8채널 입력을 사용하는 최종 구조를 채택하였다.

인코더. 이중 합성곱 블록(Conv 3×3 → InstanceNorm → LeakyReLU, $\times2$)과 2×2 최대 풀링을 4회 반복하여 공간 해상도를 단계적으로 축소한다. 각 블록의 출력 특징 맵 $\mathbf{e}_0$–$\mathbf{e}_3$은 채널 수 64, 128, 256, 512에 대해 공간 크기 128^2, 64^2, 32^2, 16^2를 가지며, 병목 블록은 8^2 해상도에서 채널 수를 1024로 확장한다.

디코더. 디코더는 두 단계의 업샘플링만 수행하여 해상도를 8^2에서 32^2로 복원한다. 각 단계에서 전치 합성곱으로 공간 크기를 두 배로 늘린 후, 대응하는 인코더 특징 맵과 채널 방향으로 연결하는 스킵 연결을 적용한다($\mathbf{e}_3$: 16^2, $\mathbf{e}_2$: 32^2). 출력 해상도보다 세밀한 $\mathbf{e}_0$, $\mathbf{e}_1$은 디코더에 연결하지 않는다.

잔차 경로. 변형 입력의 처음 4개 채널에 $4\times$ 평균 풀링을 적용한 값을 1×1 합성곱 출력에 더한 후 $[0,1]$로 클램핑하여 최종 출력을 구한다. 이 잔차 경로는, 참조 이미지를 추가 입력으로 제공하는 설계와 함께, 네트워크가 원본 픽셀을 불필요하게 변경하지 않고 변형 아티팩트에 국한된 보정만 수행하도록 구조 수준에서 강제한다.

3.3.3 손실 함수

네트워크는 다음의 복합 손실로 학습한다.

$$\mathcal{L} = \mathcal{L}_1 + \lambda_g \, \mathcal{L}_{\text{grad}} + \lambda_f \, \mathcal{L}_{\text{freq}},$$

여기서 λ_g=2.0, λ_f=0.05이며, $\mathcal{L}_1$은 픽셀 단위 평균 절대 오차(Mean Absolute Error, MAE)이다.

$\mathcal{L}_{\text{grad}}$는 인접 픽셀 차분으로 구한 수평 · 수직 기울기의 $\mathcal{L}_1$ 차이로, 픽셀 오차만으로는 포착하기 어려운 경계의 선명도를 강제한다.

$$\mathcal{L}_{\text{grad}} = \|\nabla_x p - \nabla_x t\|_1 + \|\nabla_y p - \nabla_y t\|_1$$

픽셀 공간의 $\mathcal{L}_1$만으로는 흐릿한 출력을 충분히 억제하기 어렵다. 흐림(blur)으로 인한 오차는 여러 픽셀에 분산되어 개별 픽셀 오차는 작게 유지되는 반면, 주파수 도메인에서는 선명한 경계에 해당하는 고주파 계수의 에너지가 뚜렷하게 감소한다. $\mathcal{L}_{\text{freq}}$는 이 점을 직접 이용하여, 2차원 실수 FFT 계수의 차이에 정규화된 주파수 반경을 가중치로 곱함으로써 고주파 성분의 오차를 강조한다.

$$\mathcal{L}_{\text{freq}} = \frac{1}{|\Omega|} \sum_{(u,v)\in\Omega} w_{u,v} \left| \mathcal{F}(p)_{u,v} - \mathcal{F}(t)_{u,v} \right|, \quad w_{u,v} = 1 + \hat{r}_{u,v}$$

여기서 $\hat{r}_{u,v} = r_{u,v}/r_{\max} \in [0,1]$은 정규화된 주파수 반경이다. 가중치 w는 주파수 반경에 따라 선형 증가하여, 픽셀 아트 특유의 선명한 윤곽에 해당하는 고주파 성분에 최대 2배의 패널티를 부여한다.

3.3.4 학습 절차

총 2,832쌍의 합성 학습 데이터를 사용하였으며, 배치 크기 64에서 학습률 2×10^{-4}의 Adam 옵티마이저와 코사인 어닐링 스케줄러(cosine annealing scheduler, $T_{\max} = 10{,}000$)를 적용하여 학습하였다. 데이터 증강으로는 회전($\pm30°$)과 수평 뒤집기를 사용하였다. 학습은 최대 10,000 에포크까지 수행하되, 검증 손실이 1,000 에포크 동안 개선되지 않으면 조기 종료하였다.

4 실험 결과

4.1 실험 환경 및 평가 지표

학습에는 Intel Core Ultra 9 285K (3.7 GHz), NVIDIA GeForce RTX 5090(32 GB), RAM 64 GB 환경을 사용하였으며 복원 품질 평가에는 PSNR과 SSIM [14] 두 가지 지표를 사용하였다.

PSNR. PSNR은 픽셀 단위 복원 오차를 측정하는 지표로, 알파 채널 기준 전경 픽셀만을 대상으로 RGB 3채널 평균 제곱 오차

Table 1: Synthetic-data validation results (504 pairs)

Model	PSNR (dB)	SSIM
Restore	**44.68**	**0.9986**
Restore-R15	37.93	0.9974
Restore-R30	36.96	0.9969
deformed baseline	22.19	0.9354
ref copy baseline	13.56	0.5896

(MSE, Mean Squared Error)를 계산한다.

$$\text{PSNR} = 10 \log_{10} \frac{1}{\text{MSE}}, \quad \text{MSE} = \frac{1}{N} \sum_{i} (p_i - t_i)^2$$

여기서 p_i, t_i는 [0, 1]로 정규화된 예측값과 정답 픽셀, N은 유효 전경 픽셀 수이다. 단위는 dB이며 값이 클수록 좋다.

SSIM. SSIM은 밝기(luminance), 대비(contrast), 구조(structure) 세 요소를 동시에 측정하는 지표로, 인간 지각에 더 가까운 품질을 반영한다.

$$\text{SSIM}(p, t) = \frac{(2\mu_p\mu_t + c_1)(2\sigma_{pt} + c_2)}{(\mu_p^2 + \mu_t^2 + c_1)(\sigma_p^2 + \sigma_t^2 + c_2)}$$

범위는 [0, 1]이며 값이 클수록 좋다. 본 실험에서는 32×32 RGB 전체 이미지에 대해 윈도우 크기 7로 계산하였다.

비교 기준선으로 (1) ARAP 변형 이미지를 4배 평균 풀링으로 단순 다운스케일한 *deformed baseline*과, (2) 참조 이미지를 그대로 다운스케일한 *ref copy baseline*을 설정하였다. 두 기준선은 각각 포즈 변형 자체의 한계와 포즈를 무시한 경우의 하한을 나타낸다.

평가는 두 단계로 구성된다. 먼저 합성 학습 데이터의 검증 세트 504쌍에서 네트워크의 합성 데이터 일반화 능력을 확인하고, 이어서 수동 제작한 실제 ARAP 변형 스프라이트 20쌍(세션 s01–s10, 포즈 2종)에서 실데이터 적용 가능성을 평가한다. 합성 검증 세트의 열화(warp 및 hole)는 학습 시 데이터 증강을 위해 매 반복마다 무작위로 적용되며, 평가 시에도 동일한 방식이 유지된다. 평가 수치의 재현성을 보장하기 위해 평가 시에는 난수 시드를 고정하였다.

4.2 합성 데이터 평가

Table 1과 Figure 4에 합성 데이터 검증 세트(504쌍)에 대한 정량 결과를 정리하였다. 비교 대상은 회전 증강을 적용하지 않은 *Restore*, ±15° 회전 증강을 적용한 *Restore-R15*, ±30° 회전 증강을 적용한 *Restore-R30* 세 가지 모델이다.

세 모델 모두 PSNR과 SSIM 양 지표에서 deformed baseline(22.19 dB, SSIM 0.9354)을 크게 상회하였다. 특히 SSIM 기준으로는 세 모델 모두 0.996 이상을 달성하여, 픽셀 오차뿐 아니라 구조적 유사도 면에서도 deformed baseline 대비 우수한 복원 품질을 보였다.

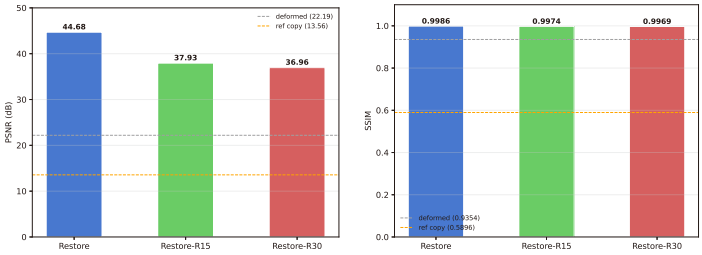


Figure 4: Validation PSNR (left) and SSIM (right) on the synthetic validation set (504 pairs) for Restore, Restore-R15, and Restore-R30. Dashed lines denote the deformed baseline (gray) and ref copy baseline (orange); higher values indicate better restoration quality.

Table 2: ARAP real-data evaluation results (all zero-shot, 20 pairs)

Model	PSNR (dB)	SSIM
Restore	12.59	0.5901
Restore-R15	12.60	**0.5973**
Restore-R30	**12.64**	0.5949
deformed baseline	12.51	0.5949
ref copy baseline	11.79	0.5395

회전 증강이 강해질수록 PSNR과 SSIM이 함께 낮아지는 것은 복원 과제의 난이도가 높아지기 때문이며, 모델 품질의 저하로 해석해서는 안 된다. Restore(base)는 고정된 포즈 분포만을 학습하여 44.68 dB / SSIM 0.9986에 달하는 수치를 달성하였으나, 이는 회전 변형 없이 고정된 포즈 패턴만을 학습한 결과로, 다양한 포즈에 대한 일반화 능력을 반영하지는 않는다. 반면 Restore-R30는 ±30° 범위의 다양한 기울기에 대한 복원 능력을 확보하는 더 어려운 과제를 수행한다.

4.3 실제 ARAP 변형 데이터 평가

Table 2는 실제 ARAP 변형 스프라이트 20쌍(s01–s10, 포즈 2종)에 대한 결과이다. 복원 네트워크는 실제 ARAP 변형 이미지를 학습에 사용한 적이 없을 뿐 아니라, 테스트에 사용된 레퍼런스 스프라이트 시트와 그 팔레트 변형본 모두 학습 데이터에서 제외되었으므로, 20쌍 전체가 완전한 제로샷(zero-shot) 조건이다.

PSNR 기준으로는 세 모델 모두 deformed baseline(12.51 dB)을 상회하여(12.59–12.64 dB), ref copy < deformed < pred의 계층 구조가 확인되었다. SSIM 기준으로는 Restore-R15(0.5973)가 deformed(0.5950)를 소폭 상회하였으며, Restore-R30(0.5949)는 deformed와 동등한 수준을 보였다. Restore(base)의 SSIM(0.5901)이 deformed를 소폭 하회하는 것은 회전 증강 없이 학습되어 ARAP 변형 특유의 기울기에 대한 대응 능력이 부족하기 때문으로 해석된다.

Table 3에서 세션별 평균 PSNR을 보면 10.4–14.8 dB에 이르는 상당한 분산이 관찰된다. s04, s06, s07 세션에서는 세 모델 모두 deformed baseline을 일관되게 상회하는 반면, s01, s03처럼 deformed baseline의 PSNR이 상대적으로 높은 세션에서는 수치

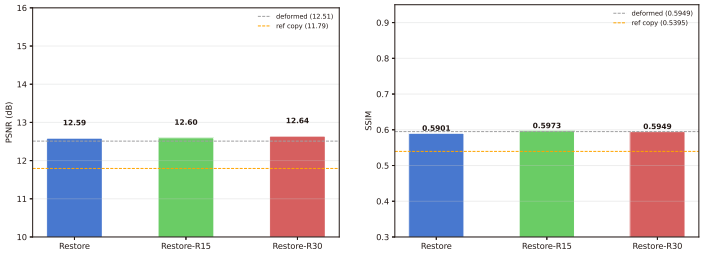


Figure 5: ARAP real-data PSNR (left) and SSIM (right) for Restore, Restore-R15, and Restore-R30 under zero-shot evaluation (20 pairs from s01–s10, two poses each). Higher values indicate better restoration quality.

Table 3: Per-session average PSNR (dB) on ARAP real-data (pose1+pose2 mean).

Session	Deformed	Restore	Restore-R15	Restore-R30
s01	13.25	11.87	12.03	11.88
s02	10.48	**10.80**	10.51	10.44
s03	15.45	14.59	14.75	14.81
s04	13.10	13.98	14.04	**14.14**
s05	12.82	12.65	12.74	**12.85**
s06	11.69	**12.38**	12.09	12.10
s07	11.40	11.69	**11.76**	11.76
s08	11.92	11.57	11.83	**12.09**
s09	11.79	**11.84**	11.79	11.81
s10	14.26	14.50	14.44	**14.53**

상 기준선을 하회하는 경우도 있다. 그러나 정성적 결과(Figure 7)에서는 해당 세션에서도 모델 출력이 시각적으로 더 우수한 복원을 보이며, 이는 PSNR이 픽셀 아트의 지각적 품질을 완전히 반영하지 못할 수 있음을 시사한다.

4.4 자명해(trivial solution) 분석

모델이 참조 이미지를 그대로 복사하는 자명해(trivial solution)에 빠지는지 여부를 확인하기 위해, alpha > 0인 전경 픽셀 각각에 대해 RGB 3채널의 최대 절대오차(ℓ_∞)를 다음과 같이 계산한다.

$$d_{\text{ref}}^{(i)} = \max_{c\in\{R,G,B\}} |p_c^{(i)} - s_c^{(i)}|, \quad d_{\text{tgt}}^{(i)} = \max_{c\in\{R,G,B\}} |p_c^{(i)} - t_c^{(i)}|$$

여기서 $p_c^{(i)}$, $s_c^{(i)}$, $t_c^{(i)}$는 각각 예측값, 참조 이미지, 정답 이미지의 채널 c 픽셀값이고, N은 유효 전경 픽셀 수이다. 임계값 $\varepsilon = 1/255$ (8비트 기준 1단계 오차)를 기준으로 두 비율을 산출한다.

$$\text{pred}{\approx}\text{ref} = \frac{|\{i : d_{\text{ref}}^{(i)} \le \varepsilon\}|}{N}, \quad \text{pred}{\approx}\text{target} = \frac{|\{i : d_{\text{tgt}}^{(i)} \le \varepsilon\}|}{N}$$

Table 4과 Figure 6에 결과를 정리하였다.

회전 증강이 강화될수록 pred≈ref 비율이 35.7%에서 26.8%로 낮아져 참조 이미지 복사 경향이 억제됨을 확인하였다. pred≈target이 함께 낮아지는 것은 증강으로 인해 복원 과제 자체가 어려워진 결과로, 단순히 정답 일치율이 떨어진다고 성능이 저하된 것은 아니다. Restore(base)의 높은 PSNR은 참조 이미지를 그대로 출력하는 경향에 상당 부분 기인하며, Restore-R30는 더 어려운 변형을 처리하면서도 자명해 경향을 억제한다는 점에서 더 나은 일반화 능력을 보인다.

Table 4: Trivial solution analysis on synthetic validation set (504 pairs)

Model	pred≈ref (%)	pred≈target (%)
Restore	35.7	83.6
Restore-R15	28.4	57.3
Restore-R30	**26.8**	53.5

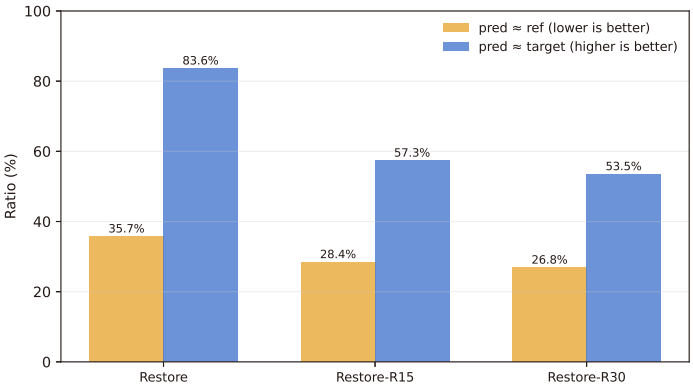


Figure 6: Trivial-solution analysis on the synthetic validation set: fraction of foreground pixels where the model output matches the reference (ref) or target image within an ℓ_∞ RGB threshold of $1/255$. Lower pred≈ref indicates weaker reference-copy behavior, while higher pred≈target indicates better target reconstruction.

4.5 시각적 결과

Figure 7은 실제 ARAP 변형 스프라이트 일부 세션(s01–s07, s10)에 대한 정성적 복원 결과를 나타낸다(지면 관계상 s08, s09는 제외하였음). 모든 이미지는 32×32 픽셀 아트를 6배 확대하여 표시하였으며, 투명 픽셀은 체커보드 배경으로 나타내었다.

세션별로 변형 난이도가 다름에도 세 모델 모두 deformed 입력의 블러링과 이미지 결손을 복원하는 경향을 보인다.

4.6 분포 외 입력에 대한 정성적 분석

Figure 8은 학습 데이터 분포에서 크게 벗어난 디자이너 제작 입력에 대한 세 모델의 출력을 비교한 것이다. 해당 입력은 정답 이미지가 존재하지 않으므로 정량 평가가 불가능하며, 출력의 시각적 품질로 모델 간 차이를 확인한다. 이는 회전 증강을 도입한 주요 동기에 해당하는 실험이다.

Restore(base)는 학습 분포와 다른 기울기나 팔다리 위치 변화가 있을 때 참조 이미지의 픽셀을 그대로 출력하는 경향이 관찰된다. 반면 Restore-R15와 Restore-R30는 동일한 입력에서 변형에

맞는 포즈 복원을 시도하며, 특히 기울어진 몸체나 뻗은 팔처럼 학습 시 거의 등장하지 않는 형태에서 더 나은 결과를 보인다. 예를 들어 Figure 8의 둘째 줄 사례에서는 Restore(base)가 팔 형태를 제대로 복원하지 못하는 반면, Restore-R15와 Restore-R30는 팔의 방향과 형태를 비교적 안정적으로 복원한다. 이는 회전 증강이 포즈 다양성에 대한 모델의 강건성을 높임을 보여준다.

5 결론

본 논문에서는 ARAP 기반 기하학적 변형과 딥러닝 복원 네트워크를 결합하여 게임 스프라이트를 생성하는 방법을 제안하였다. 합성 학습 데이터에서 모든 모델이 deformed baseline 대비 PSNR과 SSIM 양 지표에서 큰 폭의 향상을 달성하였으며, 실제 ARAP 변형 데이터의 제로샷(zero-shot) 평가에서도 ref copy $<$ deformed $\leq$ pred의 관계가 일관되게 유지됨을 확인하였다. 또한 회전 증강은 참조 이미지 복사 경향(pred$\approx$ref)을 35.7%에서 26.8%로 억제하였으며, 학습 분포를 벗어난 디자이너 제작 입력에서도 포즈 복원 강건성이 향상됨을 정성적으로 확인하였다.

한편 합성 학습 데이터와 실제 ARAP 변형 사이의 도메인 갭은 여전히 본 방법의 주요 한계로 남아 있다. 향후에는 실제 ARAP 변형 쌍의 수집 규모를 확대하고, 제안한 구성요소에 대한 추가 ablation study와 포즈 제어 사용성에 대한 보다 체계적인 평가를 수행할 계획이다.

References

[1] L. Hu, "Animate Anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8153–8163.

[2] C.-A. Hsieh, J. Zhang, and A. Yan, "Sprite Sheet Diffusion: Generate game character for animation," *arXiv preprint arXiv:2412.03685*, 2024.

[3] A. Hornung, E. Dekkers, and L. Kobbelt, "Character animation from 2D pictures and 3D motion data," *ACM Transactions on Graphics (ToG)*, vol. 26, no. 1, pp. 1–es, 2007.

[4] T. Igarashi, T. Moscovich, and J. F. Hughes, "As-rigid-as-possible shape manipulation," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 1134–1141, July 2005.

[5] M.-G. Heo and C.-H. Park, "Generating sprite animation using generative AI," *Journal of Korea Game Society*, vol. 25, no. 3, pp. 49–60, 2025.

[6] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=Fx2SbBgcte

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[8] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

[9] S. Chen and M. Zwicker, "Transfer learning for pose estimation of illustrated characters," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 793–802.

[10] Z. Yang, A. Zeng, C. Yuan, and Y. Li, "Effective whole-body pose estimation with two-stages distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2023, pp. 4210–4220.

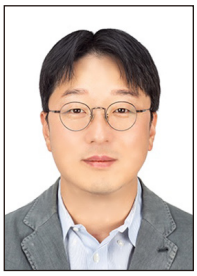
[11] E. N. Mortensen and W. A. Barrett, "Intelligent scissors for image composition," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995, pp. 191–198.

[12] PIPOYA, "PIPOYA FREE RPG Character Sprites 32x32," https://pipoya.itch.io/pipoya-free-rpg-character-sprites-32x32, accessed: 2026-05-15.

[13] ——, "PIPOYA FREE RPG Character Sprites NekoNin," https://pipoya.itch.io/pipoya-free-rpg-character-sprites-nekonin, accessed: 2026-05-15.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

〈저자소개〉



오영택

- 2006년 서울대학교 화학생물공학부, 컴퓨터공학부 학사
- 2009년 서울대학교 전기컴퓨터공학부 석사
- 2011년 서울대학교 전기컴퓨터공학부 박사
- 2011년-2015년 삼성전자 종합기술원 전문연구원
- 2015년-2016년 삼성전자 의료기기사업부 책임연구원
- 2016년-2018년 (주)알체라 CTO/공동창업자
- 2018년-2020년 (주)매그니스 연구소장
- 2021년-2025년 셀루소프트(주) 대표이사/창업자
- 2025년-현재 호서대학교 게임소프트웨어학과 조교수
- 관심분야: 3D 모델링, 증강현실, 딥러닝, 게임 그래픽스, 게임 인공지능
- https://orcid.org/0009-0003-8954-0627