Learning from Synthentic Humans

Bhuvan Channagiri, Vaibhav Kejriwal

Electrical and Computer Engineering @Northeastern University

Dec 3rd, 2025

Paper Information

- Authors: Gul Varol,Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, Cordelia Schmid
- Author affiliations : Inria, Max Planck Institute for Intelligent Systems, Body Labs Inc.
- Conference: IEEE CVPR 2017, Honolulu, HI, USA
- Number of citations: 1267 (via Google Scholar as of 12/03/2025)
- Number of GitHub repo (official or unofficial) stars: 607 (as of 12/03/2025)

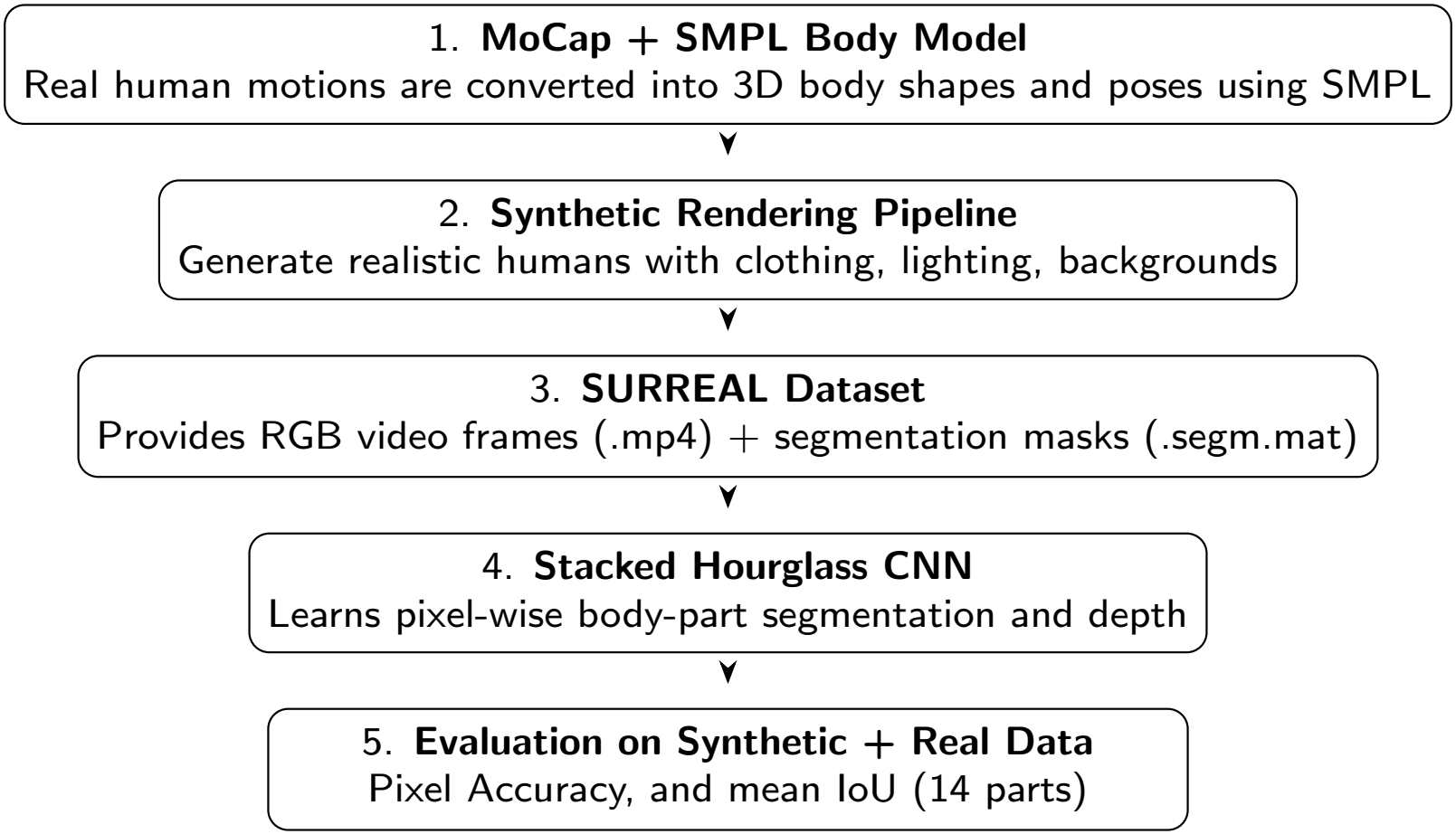
Problem

- **Original Paper Problem:**
  - Human body-part segmentation requires dense pixel-level labels, which are expensive and slow to annotate in real datasets.
  - Synthetic datasets like SURREAL provide perfect ground-truth segmentation, but models must still learn to generalize to diverse poses, textures, and lighting.
  - The original paper trains a stacked hourglass CNN **from scratch** on millions of synthetic frames, making performance heavily dependent on large-scale data.
  - No small-data strategy is explored, and segmentation quality on rare body parts or boundary regions remains limited.
- **Our Focused Problem:**
  - Can a modern transformer-based model (SegFormer-B2), combined with targeted fine-tuning, achieve strong part segmentation **even when trained on only a small subset** of SURREAL?
  - Can we improve mIoU and pixel accuracy over the original hourglass approach without relying on the full 6.5M-frame dataset?

Data: SURREAL Dataset

- **SURREAL (Synthetic hUmans foR REAL tasks)**
  - ∼6.5M frames, 67,582 clips from 145 subjects (115 train, 30 test).
  - Each rendered frame provides rich ground truth: RGB, 2D/3D joints, 14-part segmentation masks, depth maps, surface normals, optical flow, and camera/lighting parameters.
- **Subset Used in Our Work**
  - We focus exclusively on the **segmentation task** from the original paper, using only RGB frames from mp4 videos and part segmentation masks.
  - Depth estimation data is part of SURREAL, but we do not use it due to our project scope and computational constraints.
  - Due to cluster limits we train on a curated subset of SURREAL; despite using fewer samples and focusing only on segmentation, our pipeline still achieves improved mIoU and pixel accuracy.

Original Pipeline Overview



- SURREAL uses MoCap-driven synthetic humans to generate RGB frames and perfect segmentation masks, which train the hourglass CNN for pixel-wise part prediction. The model is then evaluated using pixel accuracy, and mean IoU.

Challenges in Reproducing the Original Implementation

- **Legacy Torch7 / LuaJIT Framework**
  - Original codebase relies on `Torch7`, `LuaJIT`, and older CUDA libraries.
  - Many required modules (e.g., `cunn`, `cudnn`, `inn`, `nngraph`) are no longer maintained.
- **CUDA and cuDNN Version Incompatibility**
  - Original repo requires CUDA versions in the **8.0–10.x** range and older cuDNN releases.
  - Our compute environments (Northeastern cluster & Google Colab) only support modern CUDA versions (**12.0+**), causing unavoidable build failures.
  - CMake builds for Torch7 GPU backends and LuaJIT bindings fail due to deprecated APIs.
- **Practical Implications**
  - We were unable to run or fine-tune the original hourglass Torch7 model due to irreconcilable dependencies.
  - This motivated our shift to a **modern PyTorch-based pipeline** using SegFormer-B2, which ensures stable training on current CUDA and GPU hardware.

Original Contribution: Why Change the Model?

- **Focus of our extension**
  - Instead of modifying data, losses, or evaluation, we target the **core model** in the pipeline.
  - The stacked hourglass CNN is not the most modern approach for capturing global context and fine part boundaries and hence yielding lower accuracy.
- **Why architecture matters most here**
  - SURREAL already provides rich labels (per-pixel segmentation and depth); data is not the limiting factor.
  - A better architecture should directly improve mIoU and pixel accuracy **without changing the dataset or metrics**.
- **Our design choice**
  - Replace the stacked hourglass with **SegFormer-B2**, a transformer-based segmentation model that is:
    - Strong at long-range spatial reasoning (occlusions, complex poses).
    - Compact enough to avoid overfitting in our small-data setting.

Original Contribution: SegFormer-B2 for SURREAL Segmentation

- **Segmentation-only reformulation**
  - We focus solely on **human part segmentation**, not depth, to fully exploit the segmentation labels in SURREAL.
- **Task-specific fine-tuning of SegFormer-B2**
  - Initialize SegFormer-B2 from a generic pretrained checkpoint and **fine-tune it specifically on SURREAL part segmentation**.
  - The original paper does not explore transformer-based segmentation models or such targeted fine-tuning for this task.
- **Expected and observed impact**
  - Better global context and cleaner boundaries $\Rightarrow$ higher mIoU and pixel accuracy.
  - Our results show improved segmentation quality **even with fewer training samples**.

Our Fine-Tuning Method: What We Did and Why

- **Starting Point**
  - SegFormer-B2 pretrained on ADE20K; epochs 1–6 trained on SURREAL segmentation (up to 700k samples).
  - This learns good global structure but underfits rare parts due to strong class imbalance.
- **Fine-Tuning Phase (Epochs 7–9)**
  - Loaded checkpoint from epoch 6 and reduced learning rate to 10^{-4}.
  - Applied **class-balanced cross-entropy** using inverse-$\sqrt{\text{freq}}$ weights.
  - Added **label smoothing** to stabilize boundaries and reduce over-confidence.
  - Trained on curated subsets:

    E7: 250k/10k, E8: 600k/12k, E9: 700k/14k (train/val).
- **Effect of Fine-Tuning**
  - Corrects systematic under-segmentation of small parts.
  - Maintains pixel accuracy (already dominated by background) but significantly raises mIoU.
  - Best mIoU improves from **0.401 → 0.502 for Freiburg** (+10.1 points, **+25% relative**).

Training Strategy: Ours vs Original Paper

Original Paper (Segmentation Training)

- 50K pre-training iterations on 5.3 million frames for training and 1.19 million for testing(batch size 6, LR 10^{-3}).
- No progressive staged training or fine-tuning within SURREAL.

Key Differences

- Original model uses no pretrained backbone.
- No targeted improvement for rare-part segmentation.
- Much higher data + compute requirements.

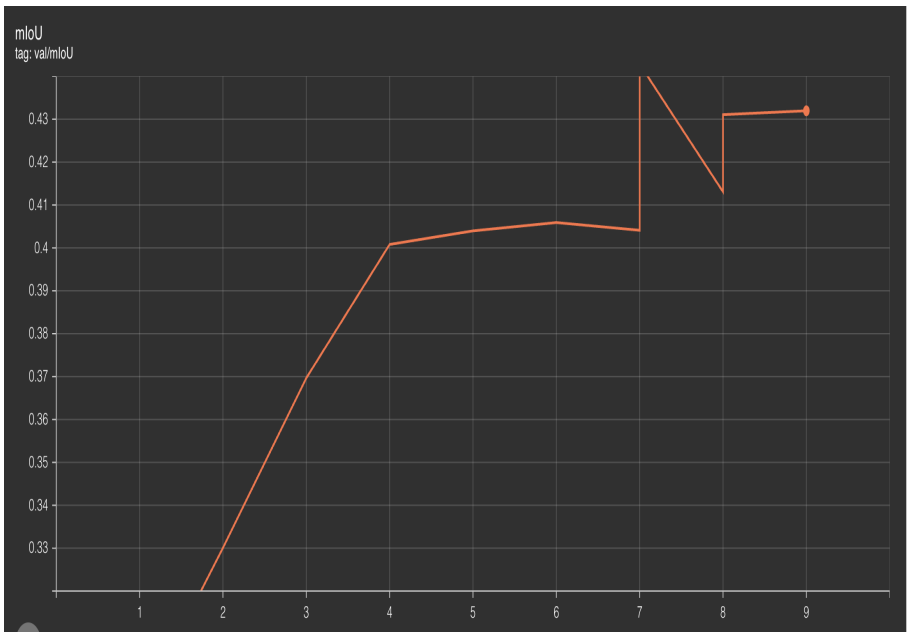
Our Training & Fine-Tuning (SegFormer-B2)

Epoch	Train	Val	Notes
1	50k	5k	Initial learning
3	200k	20k	Clear progress
4	600k	44k	Approaching saturation
6	700k	44k	Best baseline before FT
7^{FT}	250k	10k	Fine-tuning w/ lower LR
8^{FT}	600k	12k	Larger subset, stable FT
9^{FT}	700k	14k	Convergence plateau

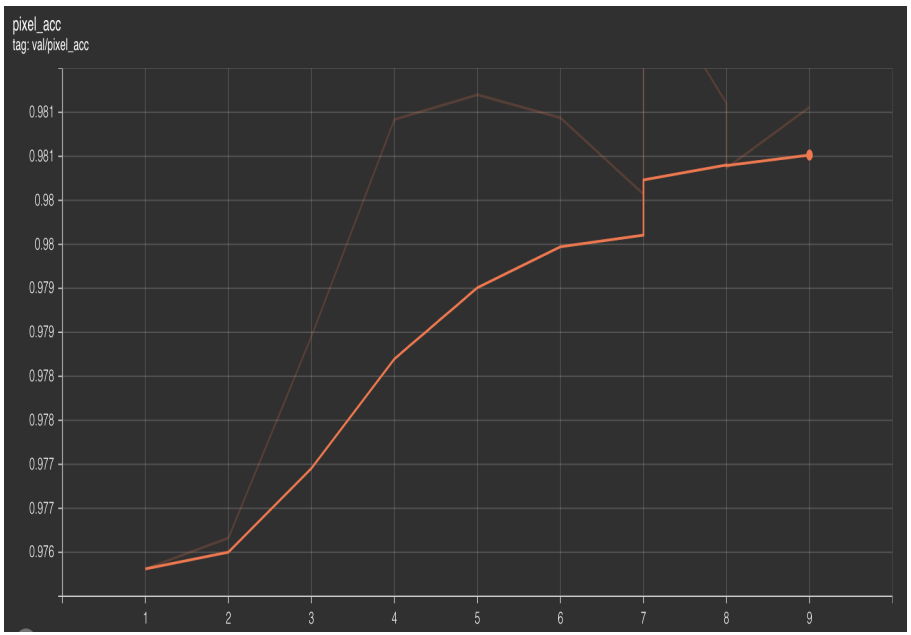
FT = fine-tuning using lower LR, class-balanced loss, label smoothing.

Results: Training Dynamics

- We monitor training and validation metrics over epochs to ensure stable convergence.
- SegFormer-B2 shows smoother improvement in mIoU and pixel accuracy compared to the Hourglass baseline.



mIoU vs. Epochs



Pixel Accuracy vs. Epochs

Results: Quantitative Evaluation on Real Dataset

- We report pixel accuracy, and mean IoU over 14 parts.
- Comparison between the Hourglass baseline and our SegFormer-B2 model.

Model	**Split**	**Pixel Acc.**	**mIoU**
Hourglass (paper)	Synthetic test	51.88%	40.10%
Segformer-B2	Synthetic subset	89.52%	48.10%
SegFormer-B2 (Finetuned)	Synthetic subset	**89.55%**	**50.20%**

- We have trained on the SURREAL dataset and evaluated on the **Freiburg Sitting People dataset** similar to the original paper.

Results: Qualitative Segmentation Examples - SURREAL

Input RGB

SegFormer-B2 Prediction

Input RGB

SegFormer-B2 Prediction

Results: Qualitative Segmentation Examples - Freiburg

Input Images

SegFormer-B2

Predictions

Thank You!

Questions?