Object Recognition Deliverable 3: Body and Cloth Depth Estimation

Pedro Agúndez Sheena Lang Zachary Parent
Martí Recalde Bruno Sánchez

May 26, 2025

Table of Contents

Introduction

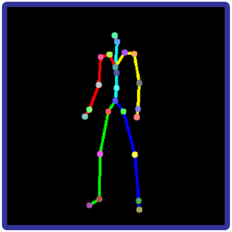
- Goal: Predict dense depth maps from monocular RGB images of clothed human subjects using the CLOTH3D dataset
- Applications: Virtual try-on, 3D avatar reconstruction, and human motion analysis
- Approach: UNet baseline with preprocessing, data augmentation, and advanced models (Attention U-Net, ResUNet-a, TransUNET)
- Enhancements: Perceptual loss via surface normals and SMPL-based pose maps for structural guidance

Preprocessing Pipeline

- Used the subset of CLOTH3D provided
- Adapted the starting-kit to return joint locations
- Generated a folder for each video, containing 256x256 images, centered on the subjects with a 10px margin
- Each folder structured in 4 subfolders: RGB, Depth, Depth visualization and Pose. Each containing an image per video frame

RGB

Depth visualization



Pose

Baseline Model

- Implemented modular 2D UNet with configurable depth, activations, and batch norm
- Trained on cropped RGB inputs to predict single-channel depth maps
- Baseline config: 5 downsampling stages, GELU activation, batch norm, $\mathrm{LR} = 1 \times 10^{-4}$
- Systematic tuning across multiple seeds and 12 training epochs per setup
- Tested variations:
  - Reduced network depth
  - No batch normalization
  - Higher / lower learning rates
  - Data augmentation (rotation, flip, color jitter, erasing, normalization)

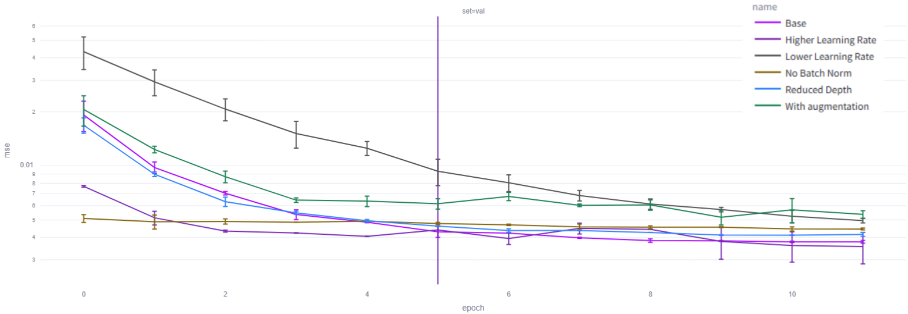
ViT Architectures

- **Attention U-net**: Enhances UNet by adding attention gates to skip connections, enabling the model to focus on relevant spatial regions and improve fine detail recovery in depth maps
- **ResUnet-a**: Integrates residual connections, dilated convolutions, and PSP pooling to capture rich multi-scale context and produce refined, normalized depth maps
- **TransUNET**:
  - Composed of three components: ResNet encoder, ViT bottleneck, and UNet-style decoder
  - Ablation study compared full pretraining (ResNet + ViT) vs. ViT-only pretraining and assessed impact on depth estimation accuracy

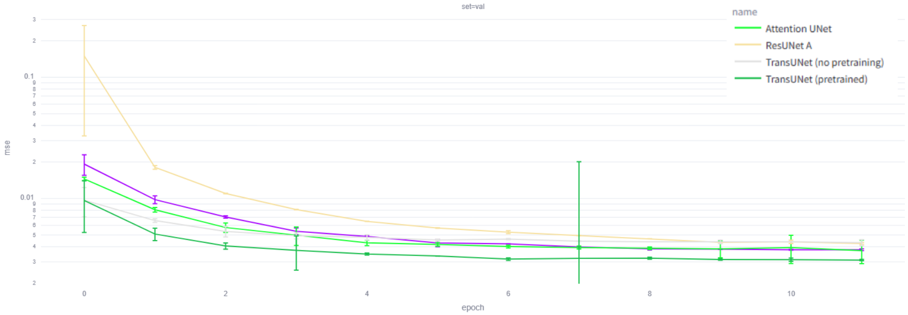
Perceptual Loss and SMPL

- **Perceptual Loss**:
  - Based on L1/L2 difference between normal maps from predicted vs. ground truth depth
  - Encourages preservation of geometric details (edges, contours)
  - Applied to TransUNet with weight 0.5
- **SMPL Pose Maps**:
  - Encodes joint positions as color-coded maps concatenated with RGB input
  - Input expanded to 6 channels
    - First ResNet layer of TransUNet adjusted accordingly
  - Aims to inject human pose priors to boost depth accuracy

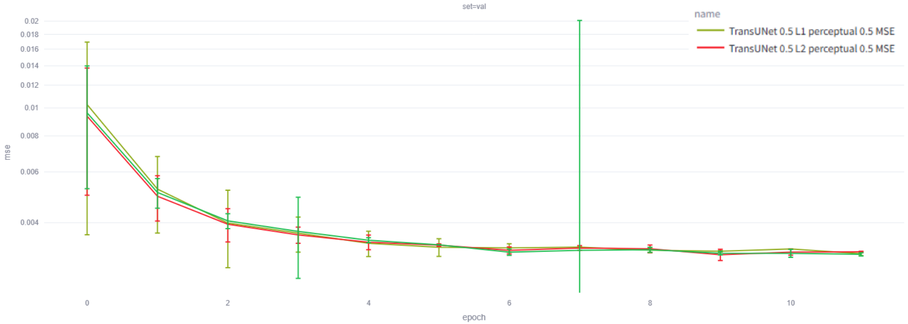
Quantitative Results: Baseline Model Tuning



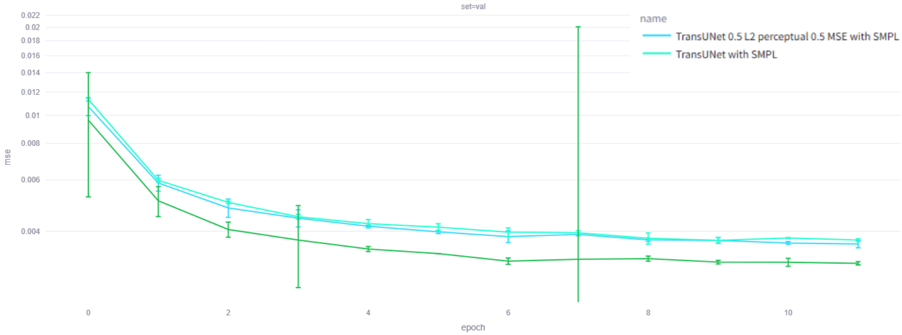
Quantitative Results: Vision Transformer Architectures



Quantitative Results: Perceptual Loss



Quantitative Results: SMPL



Qualitative Results: Inference Process

(a) Input image (b) Normalized input image (c) Ground truth (d) Depth estimation from *UNet2D base*

Figure: The inference process on a validation frame (video 158, frame 233)

.

Qualitative Results: Comparison

(a) Input image (b) Ground truth (c) *TransUNet (pretrained)* (d) *UNet2D with data augmentation*

Figure: Comparison of inference on a difficult test frame (video 173, frame 245) between the best model *TransUNet (pretrained)* and the worst model *UNet2D with data augmentation*

.

Conclusion

- **TransUNet with Pretrained Weights was the best model:**
  - Consistently outperformed UNet, Attention U-Net, and ResUNet-a.
  - Demonstrates the power of combining **transfer learning** with hybrid **CNN-Transformer architectures**.
  - Achieved a median MSE of $\mathbf{2.44 \cdot 10^{-3}}$ on the test set.
- **Advanced Features Yielded Limited Gains:**
  - Perceptual loss had minimal impact.
  - SMPL pose integration did not improve results, possibly due to disrupting pretrained weights.
- **Future Work:** Focus should be on:
  - Better integration of **pose information**.
  - Utilize **larger datasets**.

Thank you for your time!

Questions?