Free-Viewpoint RGB-D Human Performance Capture and Rendering

Phong Nguyen-Ha[1*], Nikolaos Sarafianos[2],
Christoph Lassner[2], Janne Heikkilä[1], and Tony Tung[2]

[1] Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[2] Meta Reality Labs Research, Sausalito
https://www.phongnhhn.info/HVS_Net

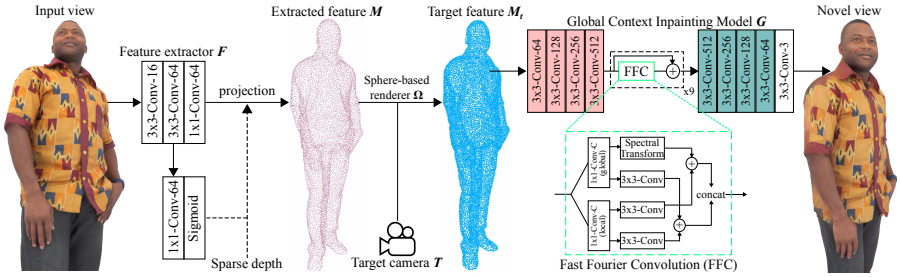

Fig. 1. *Detailed architecture of the sphere-based view synthesis network.* The feature extractor F first use three convolution layers with stride 1 to extract the features of the input view. We then infer the radius of each sphere by passing the learned features through another convolution layer and the sigmoid activation function. The green and red convolution layers of G module scale up and down the feature maps respectively.

In this supplementary material we provide additional details regarding our network designs (Sec. A), as well as implementation details (Sec. B). Additional qualitative evaluations and results are shown in the supplemental video. Finally, we discuss the limitation of our approach (Sec. C).

A Network Designs

In this section, we describe the technical details of two sub-networks of our proposed HVS-Net: a sphere-based view synthesis S and a enhancer model E.

A.1 Sphere-based view synthesis model S

Sphere-based feature warping. The architecture of the sphere-based view synthesis model S is shown in Fig. 1. Instead of directly rendering novel views using the RGB input image, we first passed it through a feature extractor F

*This work was conducted during an internship at Meta Reality Labs Research.

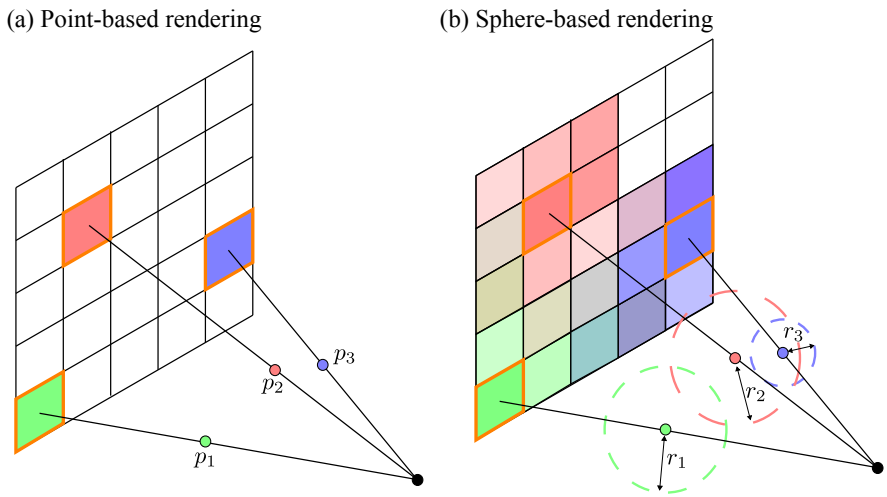

Fig. 2. Visualization of the *rendered features* between (a) point and (b) sphere-based rendering methods. Point-based method [11] can only render pixels (orange boxes) that have valid 3D coordinates. In contrast, sphere-based method [5] uses learned radius r_i of each point p_i to render neighboring pixels which leads to a denser feature map.

which consists of three convolution layers with stride 1 to maintain the spatial resolution. We choose the features f_i as the values of M where there is a valid depth value. We estimate per-sphere radius r_i by passing M to another convolution layer with sigmoid activation function. In Fig. 2, we show the visualization of rendered feature maps from a set of sparse points using point and sphere-based renderers. In case of point-based rendering [11], each 3D point p_i can render a single pixel. Therefore, a large amount of pixels can not be rendered because there is no ray connecting those pixels with valid 3D points. In contrast, the sphere-based neural renderer [5] Ω renders a pixel by blending the colors of any intersected spheres with the given ray. Since we estimate radius r_i of each sphere (dashed circle) using a shallow network, this allows us to render pixels that do not have a valid 3D coordinates. As a result, we obtain a much denser feature maps as can be seen in the Fig. 2 of the main paper. Note that, Ω is fully-differentiable and renders target feature maps very efficiently using PyTorch3D [4].

Global context inpainting model. We render the novel view using a global context inpainting model G. We design the architecture of the G module based on the encoder-decoder structure with skip connections and nine residual blocks are also utilized in the bottleneck.

In each residual block, we replace the regular convolution layers with the recently proposed Fast Fourier Convolution(FFC) [3] which possesses the non-local receptive fields. According to the spectral convolution theorem in Fourier theory, point-wise update in the spectral domain globally affects all features involved in the Fourier transform. The FFC layer splits the input features into local and

global branches. The local branch utilizes conventional convolution layers to obtain local features. In contrast, the global branch includes a Spectral Transform block [9] which uses channel-wise Fast Fourier Transform [2] to enable image-wide receptive field. The output of both branches are then summed, aggregated before adding to the residuals.

Outputs. The view synthesis model S not only predicts an RGB image I_p of the target view but also a foreground mask I_m and a confidence map I_c. We employ three different 3×3 convolution layers to predict those outputs using the output of the final layer of the G module. Thus, we apply the predicted foreground mask and confidence map to the predicted novel image as follow: $I_p = I_p \times I_m \times I_c$. We train the model S using the photometric loss $\mathcal{L}_{photo}$ as defined in the main paper.

A.2 Enhancer model *E*

Ground-truth Data: We use the RenderPeople dataset [8] to train all our models; which comprises of 1000 watertight raw meshes. To obtain IUV ground-truth we first fit an SMPL-like parametric body model to the scans and then perform non-rigid registration for all meshes and rig them for animation. In that way we obtain 1000 rigged models to which we can apply the same IUV map during rendering with an emission shader in Blender Cycles and thus obtain per-pixel perfect IUV ground-truth given an RGB input. This process is depicted in Fig. 3.

HD-IUV predictor D: Now that we have generated pairs of RGB images and ground-truth IUV maps the next step is to train a network that given an RGB image of a human, can establish accurate per-pixel correspondences **for each pixel** corresponding to the clothed human (see Fig.4). Note that the key difference between this approach and what methods such as DensePose [1] or CSE [6] are doing which is dense correspondence estimates to the unclothed human body. In addition because most approaches are trained on the DensePose-COCO dataset [1] which comprises sparse (only ~100 discrete points per human) and noisy annotations such predictions are usually inaccurate and not applicable to our application that targets clothed humans. This is also depicted in Fig. 5 of the main paper where its clear that DensePose IUV estimates result into poor texture warpings.

To train our model which we term as HD-IUV (that stands for High-Definition IUV) we employed an encoder-decoder architecture with four downsampling and upsampling convolution layers along with skip connections between them while the bottleneck comprises 3 residual blocks. This design is justified by the fact that our input-output pairs are always well aligned due to the dense correspondences established by HD-IUV which is not the case with prior work. For HD-IUV, we utilize instance normalization [10] and the ReLU activation function in all layers of the network besides the 3 output branches for each task (I, U, V outputs). The UV branches have 256 output channels (since the UV predictions can take any possible value), whereas the I channel has 25 channels which correspond to 24 body parts and background. In all branches a $1 \times$ 1 convolution is applied and

its output is an unnormalized logit that is then fed to the cross-entropy losses. Each task's scores are fed to their respective classification losses which are used to train the network as:

$$L_{IUV} = \lambda_I * L_I + \lambda_U * L_U + \lambda_V * L_V \quad (1)$$

where λ_i, L_i are the respective weighting parameters and loss functions for the I, U, V channels. Framing this problem as a multi-task learning problem (3 tasks) where the U, V and I tasks are $(256D, 256D, 25D)$ per-pixel classification problems respectively, ended up being a very effective approach to enforce strong supervisions for the surface correspondences that other losses we experimented with could not achieve. In addition we employed a silhouette loss to ensure that dense correspondence estimates are provided for each pixel of the foreground clothed human. Finally, using the predicted IUV, we can warp the occlusion-free input image to the target camera using the texture transfer technique[3] from DensePose [1].

Refinement module In this section, we utilize the warped image I_w from previous step to enhance the initially estimated target view I_p using a refinement module R. Based on the predicted confidence of the view synthesis network, we combine both images as follows: $\hat{I} = I_p + (1 - I_c) * I_w$ where $\hat{I}$ is fed to a encoder-decoder network for the refinement purposes. In this work, we try to generate humans at the novel viewpoints so rendering realistic human body parts is required. We observe that the predicted semantic I contains valuable information about the semantic information of the human in the target camera. Therefore, we use the SPADE normalization [7] to inject the semantics I to the decoder of the refinement module. As can be seen in the qualitative results, the refined image is photo-realistic compared to the ground-truth image. Note that, we use the same discriminator with [7] to perform adversarial training between both before and after refined images and the ground-truth novel views.

Discussion Here we discuss the effectiveness of our proposed HD-IUV over DensePose [1] representations to refine the target views. As can be seen in the Fig. 8 of the main paper, our Enhancer model can handle heavy occlusions using just a single photo. We emphasize that the HD-IUV representation is crucial for this refinement step because we can obtain pixel-aligned warped images at the target viewpoints compared to the ground-truth data. Therefore our warped images have higher quality compared to those produced by DensePose.

B Implementation Details

The models were trained with the Adam optimizer using a 0.004 learning rate for the discriminator, 0.001 for both the view synthesis model R and the enhancer module E and momentum parameters (0, 0.9). The input/output of our method are 1024×1024. We implement HVS-Net in PyTorch and the training across our large-scale dataset with all identities and views took 2 days to converge on 4 NVIDIA V100 GPUs.

[3] Texture Transfer Using Estimated Dense Coordinates

C Limitations

Despite producing appealing results on real-world data, the proposed method is trained solely on synthetic data. It manages to bridge the domain gap remarkably well, however we believe its performance could be further improved by integrating real-world data into the training set.

However, gathering such data is not trivial: generating (close to) noise-free point clouds for training requires elaborate multi-view capture systems, possibly enhanced with controlled lighting to simulate varying lighting conditions. A way to circumvent this partially is to train on a large-scale synthetic dataset [12] and then fine-tuning on a smaller-scale real-world dataset. This, at least, reduces the amount of data that has to be captured.

Another limitation we identified is that the warped image used as input to the enhancer model has lower quality compared to the initial estimated novel view. This is independent of the quality of the IUV mapping and is an inherent problem of the differentiable warping operation. Improving this operation could be a promising direction for future work that could increase the upper bound in quality for the novel view synthesis of fine structures in occlusion scenarios.

References

1. Alp Güler, R., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: CVPR (2018) 3, 4
2. Brigham, E.O., Morrow, R.E.: The fast fourier transform. IEEE Spectrum (1967) 3
3. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. In: NeurIPS (2020) 2
4. Johnson, J., Ravi, N., Reizenstein, J., Novotny, D., Tulsiani, S., Lassner, C., Branson, S.: Accelerating 3d deep learning with pytorch3d. In: SIGGRAPH Asia 2020 Courses (2020) 2
5. Lassner, C., Zollhofer, M.: Pulsar: Efficient sphere-based neural rendering. In: CVPR (2021) 2
6. Neverova, N., Novotny, D., Khalidov, V., Szafraniec, M., Labatut, P., Vedaldi, A.: Continuous surface embeddings. In: NeurIPS (2020) 3
7. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019) 4
8. RenderPeople: http://renderpeople.com/ 3
9. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: WACV (2022) 3
10. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016) 3
11. Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: Synsin: End-to-end view synthesis from a single image. In: CVPR (2020) 2
12. Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Johnson, M., Estellers, V., Cashman, T.J., Shotton, J.: Fake it till you make it: Face analysis in the wild using synthetic data alone. In: ICCV (2021) 5

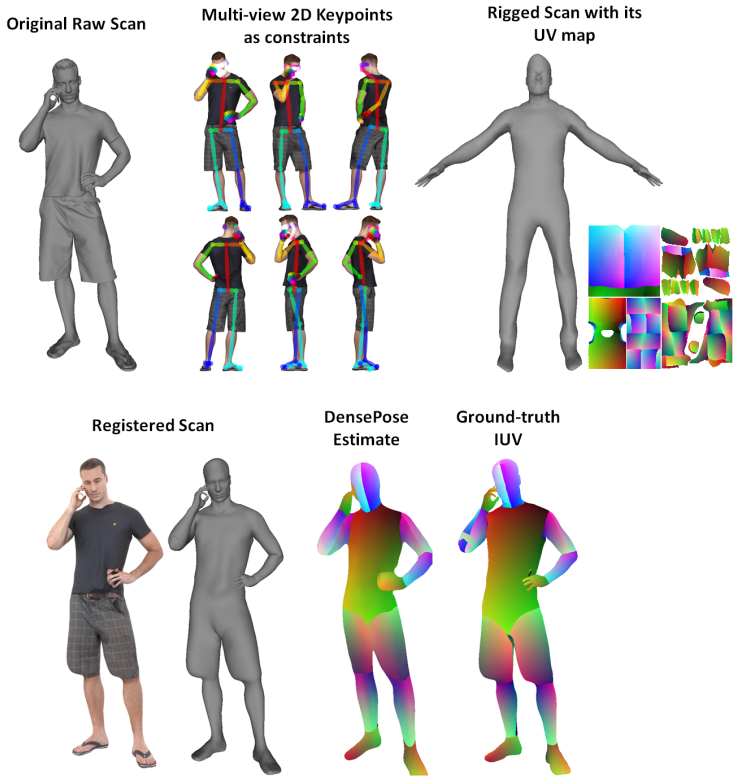

Fig. 3. *Process for IUV ground-truth generation* Given a raw synthetic scan of a clothed human (top left) we perform non-rigid registration with 2D keypoints as additional constraints (top-middle) and obtain the registered scan to the body template (bottom left) and the rigged scan (top right) which is animation ready. Using the corresponding UV map we can now obtain accurate IUV ground-truth (bottom right) that we use to train the proposed HD-IUV model. We provide the corresponding DensePose estimate to demonstrate the stark difference between the two in terms of quality as well as coverage.

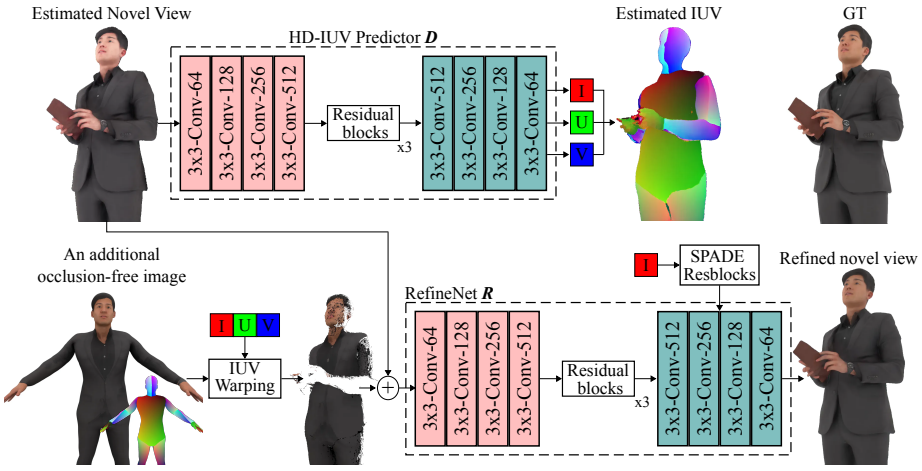

Fig. 4. *IUV-based image refinement.* Using an additional occlusion-free input, we refine the initial estimated novel view by training the Enhancer E network. We infer the dense correspondences of both predicted novel view and occlusion-free image using a novel *HD-IUV* module. The occlusion-free image is warped to the target view and then refined by an auto-encoder. The refined novel view shows better result on the occluded area compared to the initial estimated.