A. Volume Rendering Equations

We provide the mathematical derivation of the analytic form of the volume rendering equation in Equation 4 of the main article. Here we leave out the camera parameters Π by modeling in the camera coordinate. In the camera coordinate, a ray passing through pixel \((i, j)\) can be written as,

\[
r(t) = tD = t \left[ \frac{i - O_x}{f}, \frac{j - O_y}{f}, 1 \right]^T,
\]

where \((O_x, O_y)\) is the principal point of the camera and \(f\) is the focal length. Then the volume density along the ray at the \(k\)-th Gaussian kernel is,

\[
\rho_k(r(t)) = \exp\left(\frac{(t - l_k)^2}{2\sigma_k^2}\right)
\]

where

\[
l_k = \frac{M_k^T \Sigma_k^{-1} D + D^T \Sigma_k^{-1} M_k}{2D^T \Sigma_k^{-1} D},
\]

\[
q_k = -\frac{1}{2} (M_k - l_k D)^T \Sigma_k^{-1} (M_k - l_k D),
\]

\[
\sigma_k^2 = \frac{1}{D^T \Sigma_k^{-1} D}.
\]

Intuitively, \(l_k\) is the global maximizer of \(\rho_k(r(t))\) that gives the peak density of the \(k\)-th Gaussian kernel which is \(\exp(q_k)\). This enables us to calculate the integral for \(T(t)\) in Equation 4,

\[
T(t) = \exp\left(-\int_{-\infty}^{\infty} \rho(r(s)) \, ds\right)
\]

\[
= \exp\left(-\sum_{k=1}^{K} \frac{e^{q_k}}{2} \left(\text{erf}\left(\frac{t - l_k}{\sigma_k}\right) + 1\right)\right).
\]

where \(\text{erf}(x)\) is the Error Function. Then we can analytically calculate the integral of volume rendering as,

\[
\phi(r) = \int_{-\infty}^{\infty} T(t) \sum_{k=1}^{K} \rho_k(r(t)) \phi_k \, dt
\]

\[
= \sum_{k=1}^{K} T(l_k) e^{q_k} \phi_k.
\]

Here we assume the Gaussian kernels are far from the camera relative to its scale so that \(t_n \approx -\infty\) and \(t_f \approx \infty\).

B. Experiments

B.1. Evaluation on Human3.6M

We further evaluate our method on Human3.6M \cite{h36m} dataset for reference. Note that this dataset is not our main focus because 1) it does not have any occlusion; 2) it is indoor setting and the train/test data are quite similar so the performance is highly saturated (e.g., for input image of size \(224 \times 224\), assuming a \(170\)cm tall human occupies the whole image, a \(1px\) shift in the image space would correspond to \(7.6\)mm already). Here we report the performance of other SOTA methods with the same ResNet50 backbone for a fair comparison. As shown in Table 1, we achieve SOTA performance.

B.2. Comparison to Multi-modal Methods

We further compare 3DNBF with a SOTA multi-modal method \cite{6dof}, which models the conditional distribution of 3D human pose given the test image, on 3DPW–AdvOcc@80. Note that we only evaluate visible keypoints which excludes a certain amount of ambiguities. We run the official implementation and the mode prediction achieves MPJPE: 215.7 \((74.9 \uparrow)\), PA-MPJPE: 97.1 \((25.3 \uparrow)\), and PCKh: 71.7 \((13.4 \downarrow)\). The 5-sample best scores are MPJPE: 146.5 \((5.8 \uparrow)\), PA-MPJPE: 80.7 \((9.0 \uparrow)\), and PCKh: 75.8 \((9.2 \downarrow)\). This shows that our model outperforms also the multi-modal baseline. Although modeling multi-modal distributions is promising for handling severe occlusion, we consider it orthogonal to our approach.

B.3. Visualization of Image Features

We visualize the image features and pose predictions under varying occlusion levels in Fig. 1. The predicted correspondences between pixels and Gaussian kernels from the UNet features are color-coded. We can observe that the individual image features are quite robust to occlusion, only starting to get distorted when the occluder is about half the size of the human. However, as the features in the non-occluded regions are still predicted well, our method is able
<table>
<thead>
<tr>
<th>Method</th>
<th>Human3.6M [1]</th>
<th>MPJPE ↓</th>
<th>PA-MPJPE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMR [3]</td>
<td>88.0</td>
<td>56.8</td>
<td></td>
</tr>
<tr>
<td>SPIN [5]</td>
<td>62.3</td>
<td>41.7</td>
<td></td>
</tr>
<tr>
<td>HMR-EFT [2]</td>
<td>-</td>
<td>46.0</td>
<td></td>
</tr>
<tr>
<td>PARE [4]</td>
<td>82.7</td>
<td>53.7</td>
<td></td>
</tr>
<tr>
<td>3DNBF</td>
<td><strong>58.7</strong></td>
<td><strong>38.9</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Evaluation on Human3.6M protocol 2.

to correctly predict the non-occluded joints due to the robust likelihood (Eq.3)

Figure 1. Visualizing image feature and pose estimation under varying occlusion levels. Left: GT image-3D correspondence. Right top to bottom: input occluded image, PARE output, our 3DNBF output, and predicted image-3D correspondence.

**B.4. More Analysis of NBV compare to Mesh representation.**

In our ablation study, the mesh baseline is implemented with the 3D-aware features, the contrastive training, and the robust likelihood. Compared to the mesh representation with SoftRas [6], the volume representation is analytically differentiable and hence can provide better gradients, particularly in the case of self-occlusion, due to the better volume density blending compared to the distance-based blending used in SoftRas. Fig. 2 illustrates the optimization process under self-occlusion with a mesh representation and our NBV. We set $\sigma=10^{-3}$, $\gamma=10^{-2}$, and $K=40$ for SoftRas which is consistent with the parameters used in the ablation. Note how initially the right arm is estimated to be behind the body, but from iteration 40 can be corrected to be in front of the body.

**B.5. Qualitative Results**

In Fig. 4, 5 and 6, we provide more qualitative results for 3DNBF comparing with state-of-the-art 3D human pose estimation methods on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. Qualitative results on 3DOH50K are provided in Fig. 7. The comparison between our 3DNBF and optimization-based methods: EFT [2] and 3D POF [8] are visualized in Fig. 8 and 9 respectively.

**B.6. Limitations and Failure Cases**

While being robust to occlusion, our method do have some limitations, which we leave for future work. The limitations of our method are: 1) the inference speed is not real-time; 2) more detailed body models may explain the observed features better and improve accuracy; 3) it should be extended to multi-person scenarios. One of the main failure cases we observe is the occasional front-back switch errors when the head is occluded as shown in Figure 3.

**References**


Figure 4. Qualitative results on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. For left to right are the input image, (a) initial pose predicted by SPIN [5], (b) 3DNBF prediction and (c) ground truth pose.
Figure 5. Qualitative results on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. For left to right are the input image, (a) initial pose predicted by HMR-EFT [2], (b) 3DNBF prediction and (c) ground truth pose.
Figure 6. Qualitative results on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. For left to right are the input image, (a) initial pose predicted by PARE [4], (b) 3DNBF prediction and (c) ground truth pose.
Figure 7. Qualitative results on 3DOH50K [9]. For left to right are the input image, (a) initial pose predicted by regression-based methods, (b) 3DNBF prediction and (c) ground truth pose. Row 1-2, 3-4, and 5-6 are results for SPIN [5], HMR-EFT [2], and PARE [4] respectively.

Figure 8. Qualitative results on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. For left to right are the input image, (a) Optimization results of EFT [2], (b) 3DNBF prediction and (c) ground truth pose. Initial poses are predicted by HMR-EFT [2].
Figure 9. Qualitative results on 3DPW-AdvOcc@40 and 3DPW-AdvOcc@80. For left to right are the input image, (a) Optimization results of 3DPOF [8], (b) 3DNBF prediction and (c) ground truth pose. Initial poses are predicted by HMR-EFT [2].