NP-Hand: Novel Perspective Hand Image Synthesis Guided by Normals

—Supplementary Material—

I. Discussion on different guidance.

For the guidance candidates in Fig.4 of the main paper, we provide more discussion on their characteristics. Among them, values in depth maps may be ambiguous for our task. To ensure the consistency of hand poses between synthesized images, maintaining the consistency of the guidance is particularly important. However, it is very challenging to estimate consistent depth maps from source and target images respectively, thus causing ambiguity in the synthesis process. The creation of an IUV map requires accurate hand mesh and additionally depends on fixed UV unwrapping. The more cumbersome process compared to creating a normal map makes us abandon it. In addition, similar to depth maps, estimating accurate IUV maps from input images is also challenging. As for the pose map, the main reason we ignore it is that the guided information it provides is sparse. Just as we enhanced at the beginning of the introduction section, synthesizing novel-view hand images faces obstacles to the similar appearance and higher articulation of human hands. This means appropriate guidance should provide more structural information, otherwise artifact problems may occur. The first row of Fig.12 (b) in the main paper provides strong evidence for our viewpoint, where within the pose maps as guidance, the synthesized image has an opposite orientation to the target image. By comparison, a normal map circumvents most shortcomings and is regarded as guidance for synthesizing novel-view hand images.

II. Advantages of normal maps.

To further discuss the advantages of normal maps in promoting consistency between different perspectives, we introduce more supportive experiments. Specifically, we construct 3D Gaussian Splatting using multi-view images synthesized with different maps as guidance and analyze their practical effects. There is a consensus in this experiment that the better the consistency between synthesized perspectives, the higher the quality of the reconstructed 3DGS should be. To this end, we respectively train our method using various guidance, including depth maps, IUV maps, pose maps and normal maps. Subsequently, at the inference stage, we generate multi-view images guided by diverse maps from a single image as input. Detailed methodologies for obtaining these maps are outlined in the main paper. Since the absence of background constraints in our method, achieving consistency in the background among synthesized images is challenging. Therefore, we perform an *AND* operation between the synthesized images and the corresponding maps to obtain masked hand images, and further use them to construct 3DGS models with 30000 iterations.

Fig. 1. With normal maps as guidance, the constructed 3DGS could render hand images with more intricate local details and fewer artifacts.

TABLE I
Higher PSNR indicates that the synthesized images, guided by the normal map, help to construct high-quality 3DGS.

Guidance	PSNR ↑			
	FrameA	FrameB	FrameC	FrameD
Pose map	30.972	29.907	29.805	28.441
Depth map	31.372	29.450	30.003	29.362
IUV map	**32.282**	30.326	30.651	29.904
Normal map	32.020	**30.552**	**31.090**	**30.251**

The visualizations in Fig. 1 demonstrate that the 3DGS with normal maps receives superior performance, as it could render higher realistic images, featuring more intricate local details and fewer artifacts. Tab. I reports similar conclusions that employing normal maps as guidance achieves the highest PSNR values. Although IUV maps as guidance also yield realistic visual outcomes, as we discussed in the above section, not only obtaining its ground truth is more complex but also estimating it from the input images is more challenging. In summary, using normal maps as guidance has ensured the quality of synthesized images and promoted the exploration of downstream tasks such as 3DGS. In the table, FrameA refers to the frame of Capture1-ROM07_Rt_Finger_Occlusions-frame27370, FrameB refers to Capture1-ROM07_Rt_Finger_Occlusions-frame28204, FrameC refers to Capture1-ROM08_Lt_Finger_Occlusions-frame28571 and FrameD refers to Capture1-ROM08_Lt_Finger_Occlusions-frame29423.

TABLE II
QUANTITATIVE COMPARISONS ON *Interhand2.6M* AND *Hand4K++*, WITH THE SAME NORMAL MAPS AS GUIDANCE.

Methods	*Interhand2.6M*				*Hand4K++*			
	PSNR ↑	LPIPS ↓	SSIM ↑	FID ↓	PSNR ↑	LPIPS ↓	SSIM ↑	FID ↓
PATN	16.533	0.278	0.304	97.747	14.082	0.214	0.366	64.829
DPTN	16.137	0.291	0.319	102.640	15.079	0.206	0.375	68.560
GFLA	16.885	0.270	0.349	97.790	15.080	0.210	0.392	52.253
CoCosNet V2	16.754	0.260	0.375	43.252	14.891	0.239	0.389	55.586
NTED	16.949	0.233	0.381	38.671	15.143	0.208	0.401	52.547
PIDM	17.441	0.237	0.398	24.420	15.210	0.228	0.412	43.249
CFLD	17.775	0.210	0.392	26.659	**15.707**	0.200	0.419	40.551
Ours	**17.868**	**0.202**	**0.412**	**12.009**	15.644	**0.189**	**0.427**	**34.470**

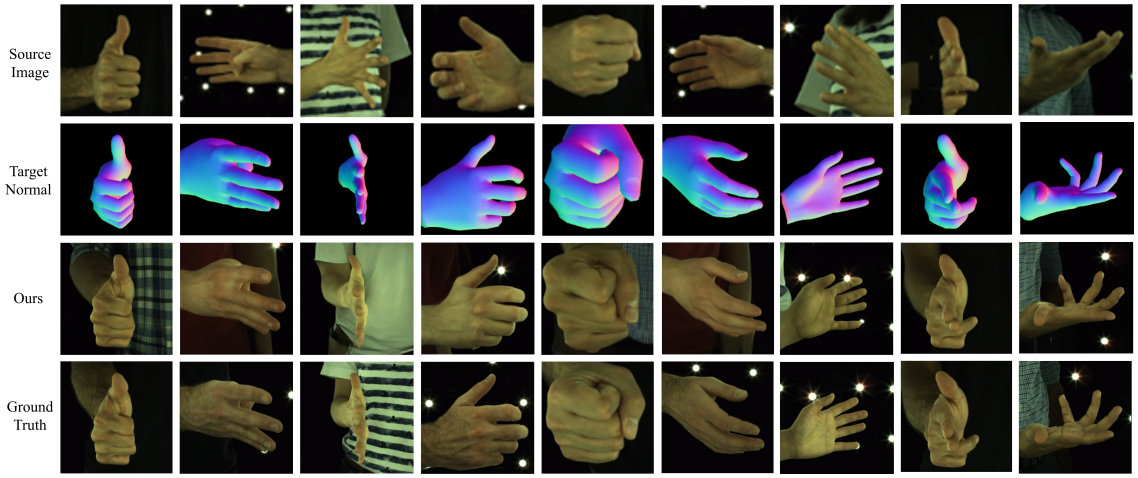

Fig. 2. More synthesis on the *Interhand2.6M* dataset.

Fig. 3. More synthesis on the *Hand4K++* dataset.

III. MORE COMPARISONS AND RESULTS.

For more fair comparisons, we have trained all methods using the same normal maps as guidance and report the quantitative results in Tab. II. We observe that all the baselines achieved improvements to varying degrees. This also indicates that using normal maps as guidance for synthesis is indeed beneficial. Nevertheless, in conjunction with Fig.9 in the main paper, we find that CoCosNet v2 still does not perform well, as the synthesized images show discrepancies in appearance compared to the source image. We attribute this to the sparse texture characteristics of human hands, as the key for CoCosNet V2 is to search cross-domain correspondences. This is more challenging for hands compared to clothed people.

We further illustrate more synthesized results from different datasets in Fig. 2 and Fig. 3.