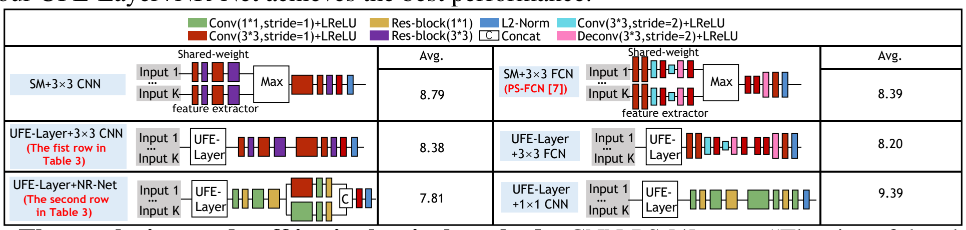
We sincerely thank all reviewers for their valuable suggestions. Below we respond to the comments and concerns.

R1&R2: Effectiveness of UFE-layer. We use the fixed network architecture to further demonstrate the effectiveness of UFE-Layer in the table below. We also add the 1×1 network suggested by R2 in the table. For convenience, we use "SM" to represent the "shared-weight feature extractor + max-pooling" strategy of PS-FCN [7]. The comparison between SM+3×3 CNN and UFE-Layer+3×3 CNN and the comparison between SM+3×3 FCN and UFE-Layer+3×3 FCN prove the effectiveness of UFE-Layer. Although the first row of Table 3 (UFE-Layer+3×3 CNN) achieves almost the same performance as PS-FCN (SM+3×3 FCN), it does not mean that UFE-Layer has no contribution compared to SM. PS-FCN uses a Fully Convolutional Network (3×3 FCN), and part of the 3×3 layers have strides as 2 to achieve down-sampling and up-sampling. We think that setting strides as 2 has a similar effect to the 1×1 layers in our NR-Net, *i.e.*, weakening the mutual influence among pixels and reducing over-smoothing in the spatial domain. As shown in the table below, our UFE-Layer+NR-Net achieves the best performance.



R1&R3&R4: The resolution trade-off in single-pixel methods. CNN-PS [4] states:"The size of the observation map (w) should be chosen carefully. As w increases, the observation map becomes sparser. On the other hand, the smaller observation map has less respresentability." The later methods based on observation maps [5, 6] also clearly indicate: When the input number decreases, it causes serious sparsity problem and performance degradation, which is an obvious manifestation of resolution trade-off. Therefore, CNN-PS only conducts experiments under dense inputs (96) by setting a large w (32). However, its performance drops significantly under sparse inputs, as shown in Table 1. LMPS [5] sets w as 14 to achieve better performance under sparse inputs, but its performance under dense inputs has declined. This clearly reflects that it is not easy to take the resolution trade-off. Although GPS-Net is not as good as CNN-PS under dense inputs, we achieve the best results when the input number is less than 64, and the best overall performance, as shown in Table 1.

R1: The suggested ablation study (Observation maps + Pooling + NR-Net). We respectfully point out that we should not fuse the feature maps generated by observation maps through any pooling operation (like PS-FCN [7]). Because the max-pooling [7] is performed for feature maps under different lightings, but the feature maps of observation maps are generated for each pixel, with varying lighting information encoded in the observation maps.

R1&R2: The ablation study of Eq. (5). Due to the limited rebuttal time, we only quickly test the effect of max-pooling and averaging using a smaller dataset. We find that max-pooling achieves better performance than averaging (also demonstrated in Section 4.1 of PS-FCN [7]), while averaging is more robust to the varying input numbers. The combination of them achieves the best results with robustness. We are conducting the complete test and will add the detailed ablation study in the final version.

R2: The ablation study of SGC filters. The general spectral graph convolution networks require the input graphs to have the same topologies, *i.e.*, the fixed input number during training and testing, which is similar to DPSN [3]. Like the state-of-the-art methods [4-7,29], we aim to handle an arbitrary number of inputs, and hence we use our SGC filters to handle graphs containing an arbitrary number of adjacent nodes (graphs with inconsistent topologies).

R2: The training number. To show the flexibility of our model, we did not use the same training number as the testing number. We have trained three models under 4, 8 and 16 inputs to test under {4}, {8,10} and {16,32,64,96} inputs, respectively.

R4: The contribution of NR-Net. We respectfully point out that GPS-Net is our entire network including UFE-Layer and NR-Net. The contribution of NR-Net is demonstrated by the comparison with "all-pixel" methods [7,29]. Qualitative results in Figure 2 (paper) and Figure 2 (supplement) illustrate that NR-Net can predict normal maps with richer details. It is quite important in photometric stereo that aims at acquiring high-resolution 3D information [1].

R4: The qualitative comparisons in Figure 2. The qualitative comparisons with state-of-the-art methods including NEURAL-PS [29] were shown in Figures 5-39 in the supplementary material. The code for LMPS [5] has not been published online. Hence, we contacted the authors, and they were only able to provide the numbers on benchmark data. Figure 2 is just an example to show the contribution of NR-Net in preserving high-resolution details compared with "all-pixel" methods [7,29]. Hence, we chose the best-performing all-pixel method PS-FCN [7] in Table 1 as a representative for comparison. Figure 2 in the supplementary material gives more results.

R1&R2&R4: Other suggestions. We will make the following changes as suggested in the final version. 1) We will separate Figure 1 and show it in a clear way (R1, R2). 2) Since DiLiGenT [2] is currently the only real-world photometric stereo benchmark, we will try our best to test GPS-Net on more synthetic data with diverse BRDFs and shapes (R2, R4). 3) We will make a runtime comparison and show how multi-scale implementation is achieved for NR-Net (R2).