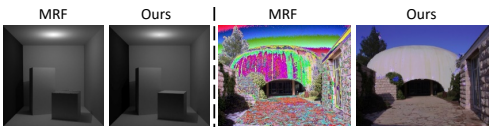
We sincerely thank all reviewers for their valuable comments and suggestions. We respectfully think it should be interesting to share with the NeurIPS community how to learn unwrapping a natural image, though the camera might be obscure. We will fix the typos and improve the figures as suggested in the final version. Below we respond to specific comments and concerns.



R1&R2: Results of the MRF algorithm [55]. Those results are indeed generated with the original implementation (we have got the code from the authors). We noticed and confirmed with the authors that the parameters of the MRF algorithm (especially the energy function and the cliques to consider) need carefully fine-tuning on different input data to achieve optimal results, and the fine-tuning process is quite tricky. Moreover, for RGB data, it independently unwraps each channel with different parameters. According to our experiments, the images shown in [55] should be "cherry-picked", as mentioned by R1. Since our method uses the same set of parameters for all test cases, for a fair comparison, we fix the parameters of the MRF algorithm for evaluation as well. As shown in the right figure (examples are from Figure 4 and 9 of [55]), by applying the MRF algorithm using the same set of parameters, the left image is well recovered while the right one is not. In comparison, our learning-based method gives consistently good results.

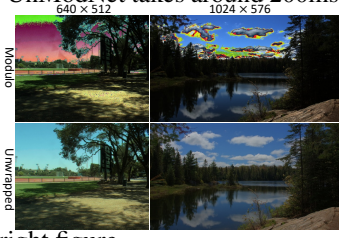
R1: Other non-learned techniques for processing modulo imagery. Please kindly note that our method only takes a *single* modulo image as input. We did not find other suitable baselines. Papers mentioned in the review "Robust Multi-Image HDR Reconstruction for the Modulo Camera", "Reconstruction from Periodic Nonlinearities, with Applications to HDR Imaging", "Signal Reconstruction from Modulo Observations", all require multiple images as input (what's more, the model presented in the first paper is limited to two modulo periods). Multiple modulo images could be merged for HDR as done in [55], but such a multi-image approach is beyond the scope of our paper. We will clarify this assumption in the final version, and consider learning to unwrap multiple modulo images as our future work.

R1: Evaluations on other conventional CNN models. We respectfully point out that we have included such results in our ablation study. For example, we have tested a vanilla attention U-Net structure that predicts the number of rollovers **K** (referring to "Directly predicting **K**" in Table 2 of the paper). To show the effectiveness of modulo edge separator, we have also tried a single rollover mask predictor to directly output binary rollover mask (referring to "Removing modulo edge separator" in Table 2 of the paper). The final rollover mask predictor adopts a modified attention U-Net structure, and it outperforms a vanilla attention U-Net structure by 2.21 dB for PSNR according to our previous experiments (when we were designing our network). We will include the results of such a vanilla attention U-Net in the final version.

R2: Other deep-learning based 2D phase unwrapping methods. We have tried our best to search for the code of such methods, *e.g.*, [41, 44, 49], but none of them is available now. We are afraid that re-implementing them within the rebuttal period could not produce reliable results. We will try to include such a comparison in the final version. However, we suspect these methods designed for phase maps instead of modulo images could not compete with either [55] or our method, due to the essentially different natures of phase maps and wrapped natural images.

R2: Details about the tone mapping method. We use the OpenCV-Python's Reinhard tone mapping algorithm (`cv2.createTonemapReinhard(intensity=-1.0, light_adapt=0.8, color_adapt=0.0)`) for visualization.

R2&R3: Implementation details of the MRF algorithm [55]. The code of [55] is written in MATLAB script and cannot run on GPU, so we run it on an Intel Core i7-8700K CPU (using a single core), while UnModNet runs on an NVIDIA 2080 Ti GPU (thanks to the PyTorch's CUDA acceleration). At each iteration, UnModNet takes around 200ms to process a 512×512 modulo image, and [55] takes around 24s.



R2: Evaluation metrics (SSIM, PSNR, MS-SSIM). We have further conducted evaluations on the tone-mapped HDR images with metrics SSIM, PSNR, and MS-SSIM, and the advantage of our method did not change. For example, the PSNR (dB) results in Table 1 of the paper become 30.33 for our method, 7.73 for MRF [55], 19.99 for DrTMO [7], 18.79 for ExpandNet [30], and 18.68 for HDRCNN [6]. We will include complete results in our final version.

R2: High-resolution results. Examples (scaled to the same height) are shown in the right figure.

R2: Physical limitations and properties of the sensor. The robustness might be improved if we include a proper camera noise model in our dataset generation pipeline, and we consider to add it in our future work. Our method could adapt to other bit depth (*e.g.*, 12-bit) if we change the bit depth of the modulo images in the training dataset.

R3: Pros and cons of different single-image HDR reconstruction methods. We will improve the the related work section by comprehensively discussing pros and cons of each method.

R4: Details of our dataset. Our training dataset consists of 21312 pairs of modulo and updated (referring to Equation (1) of the paper) image patches in 256×256 resolution, which are randomly cropped and generated from 600 HDR images collected from various sources [12, 11, 13, 14, 37, 53, 10, 27]. The details of our dataset creation pipeline and the parameter settings could be found in Section 7 of the supplementary material.

R4: Evidence of not overfitting and the generalization applicability. The evaluation on two types of real captured data could prove that our method does not overfit, and has a good generalization ability.