# Supplementary Material for WACV 2025 ID 1212
# Harmonizing Attention: Training-free Texture-aware Geometry Transfer

Eito Ikuta[1] Yohan Lee[1] Akihiro Iohara[1] Yu Saito[1] Toshiyuki Tanaka[2]

[1]DATAGRID Inc. [2]Graduate School of Informatics, Kyoto University

## 1. Comparison with Poisson Image Editing

While our primary comparative evaluations focused exclusively on deep learning-based methods, we herein extend our experimental analysis to include Poisson Image Editing (PIE) [1], a representative classical image manipulation technique. PIE is fundamentally a gradient-domain image synthesis method that synthesizes images by solving the Poisson Equation defined through Laplacian operators, thereby reconstructing image regions through gradient optimization. Specifically, we adopted the Fourier-domain approach of PIE, employing a vector field $\mathbf{v}$ derived from the Laplacian with the maximum absolute value between source and target images — a strategy referred to as the "Maximum" approach in [1]. This methodological selection was predicated on its potential to generate visually harmonious composites across diverse images.

Qualitative assessment revealed the significant limitations of PIE when compared to our proposed method. While PIE struggles to maintain color consistency and preserve source image geometry, our approach demonstrates superior image synthesis capabilities (Fig. 1). In particular, a careful examination of the first and third rows in Fig. 1 reveals the critical deficiency of PIE in preserving source geometry, with notable instances of complete occlusion or unintended elimination of complex structural elements such as cracks and apertures. These observations can be interpreted as revealing the fundamental limitations inherent in synthesizing textural information and geometry within the Laplacian domain, which fundamentally constrains the capability of gradient-based image editing techniques to preserve comprehensive image characteristics.

We extended the comparative assessment framework employed in Table 2 of the main manuscript to evaluate PIE using identical metrics (Table 1). The quantitative analysis revealed a nuanced performance profile: while PIE demonstrated exceptional similarity with background regions, it exhibited critically compromised performance in preserving source foreground geometric characteristics. Specifically, metrics assessing foreground region similarity uniformly indicated near-minimal performance, suggesting a fundamental limitation in PIE's ability to maintain the geometry of source image features during image synthesis.

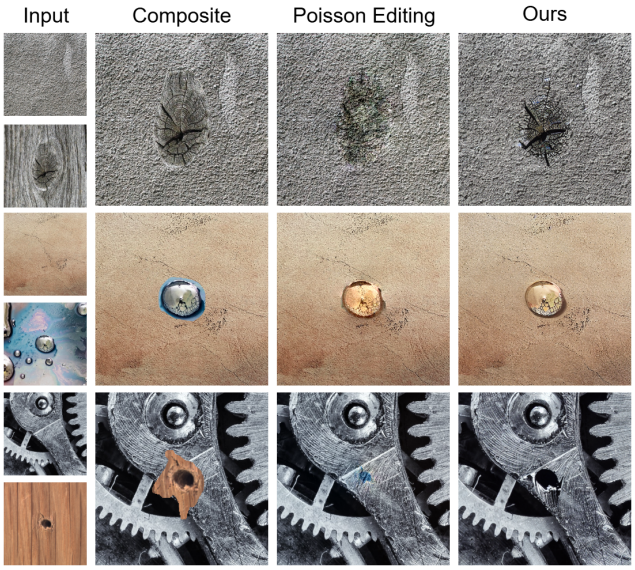


Figure 1. Additional qualitative comparison with PIE. The methods to be compared are set to maximize the quality of the generated images, and the highest quality results are selected and included. The same images as Figure 3 in the main manuscript are used except for PIE.

## 2. Details of User Study

In the user study, we showed all inputs-outputs pairs to every participant and asked him/her to score each output as described in Sec.4.4 of the main manuscript. Here, to get fair answers, all participants completed the questionnaire without knowing which image corresponded to which method.

## References

[1] J. M. Di Martino, G. Facciolo, and E. Meinhardt-Llopis, "Poisson Image Editing," *Image Processing On Line*, vol. 6, pp. 300–325, 2016. https://doi.org/10.5201/ipol.2016.163. 1, 2

Table 1. Quantitative evaluation results for geometry composition in a given target background image. Arrows next to each score indicate score interpretation: ↓ lower is better, ↑ higher is better. The minimum value for LPIPS and DISTS is 0, and the maximum value for CLIP is 100. A total of 150 images were used for the evaluation.

| Method | $\text{LPIPS}_{(\text{bg})}$ ↓ | $\text{LPIPS}_{(\text{fg})}$ ↓ | $\text{CLIP}_{(\text{bg})}$ ↑ | $\text{CLIP}_{(\text{fg})}$ ↑ | $\text{DISTS}_{(\text{bg})}$ ↓ | $\text{DISTS}_{(\text{fg})}$ ↓ |
|---|---|---|---|---|---|---|
| Poisson Image Editing [1] | **0.037** | 0.415 | **86.909** | 72.168 | **0.118** | 0.279 |
| Paint By Example [2] | 0.424 | 0.273 | 70.694 | 84.348 | 0.256 | 0.241 |
| TF-ICON [3] | 0.434 | 0.392 | 69.470 | 81.422 | 0.309 | 0.315 |
| PHDiffusion [4] | 0.408 | 0.255 | 73.624 | 84.879 | 0.276 | **0.196** |
| TF-GPH [5] | 0.324 | **0.252** | 71.234 | 86.936 | 0.223 | 0.237 |
| Ours | 0.266 | 0.255 | 68.180 | **91.352** | 0.179 | 0.250 |


[2] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen, "Paint by example: exemplar-based image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18381–18391, 2023. 2

[3] S. Lu, Y. Liu, and A. W.-K. Kong, "TF-ICON: Diffusion-based training-free cross-domain image composition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2294–2305, 2023. 2

[4] L. Lu, J. Li, J. Cao, L. Niu, and L. Zhang, "Painterly image harmonization using diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, p. 233–241, Association for Computing Machinery, 2023. 2

[5] T.-F. Hsiao, B.-K. Ruan, and H.-H. Shuai, "Training-and-prompt-free general painterly harmonization using image-wise attention sharing," 2024. arXiv:2404.12900v1 [cs.CV]. 2