Solving Vision Problems via Filtering
Supplementary Material

Sean I. Young[1] sean0@stanford.edu
Aous T. Naman[2] aous@unsw.edu.au
Bernd Girod[1] bgirod@stanford.edu
David Taubman[2] d.taubman@unsw.edu.au

[1]Stanford University [2]University of New South Wales

1. Experimental Results

In our main paper, we provided experimental results for a number of vision-related inverse problems. This supplement provides additional details on the formulations used, as well as more extensive visual results for the experiments.

1.1. Disparity Super-resolution

For our disparity super-resolution experiment, we use the dataset from [1], which is a subset of the Middlebury stereo dataset. We show visualizations of our $16\times$ super-resolution disparity maps in Figure 4.

1.2. Optical Flow Estimation

In our experiments, we use the color-gradient constancy model [2] instead of the brightness-constancy one [3]. In all cases, one can express the optical flow data fidelity term as

$$d(\mathbf{u}) = \|\mathbf{H}(\mathbf{u} - \mathbf{u}_0) + \mathbf{z}_t\|_2^2 \tag{S1}$$

see (27) in our main paper. The color-constancy model gives us

$$\mathbf{H} = \begin{bmatrix} \mathbf{Z}_x^R & \mathbf{Z}_y^R \\ \mathbf{Z}_x^G & \mathbf{Z}_y^G \\ \mathbf{Z}_x^B & \mathbf{Z}_y^B \end{bmatrix}, \quad \mathbf{z}_t = \begin{bmatrix} \mathbf{z}_t^R \\ \mathbf{z}_t^G \\ \mathbf{z}_t^B \end{bmatrix} \tag{S2}$$

in which $\mathbf{Z}_{x,y}^{R,G,B}$ denotes the x- and the y-derivatives of the target image in the R, G and B components, and $\mathbf{z}_t^{R,G,B}$ are the difference of the reference image from the target one, in the R, G and B image components.

The gradient-constancy model on the other hand gives us the derivative data

$$\mathbf{H} = \begin{bmatrix} \mathbf{Z}_{xx} & \mathbf{Z}_{xy} \\ \mathbf{Z}_{yx} & \mathbf{Z}_{yy} \end{bmatrix}, \quad \mathbf{z}_t = \begin{bmatrix} \mathbf{z}_{xt} \\ \mathbf{z}_{yt} \end{bmatrix} \tag{S3}$$

in which $\mathbf{Z}_{xx}$, $\mathbf{Z}_{xy}$ and $\mathbf{Z}_{yy}$ are the second-order derivatives of the target image, and $\mathbf{z}_{xt}$ and $\mathbf{z}_{yt}$ are the difference of the first-order differences of the reference image from the target ones. When the gradient constancy model is applied on each of the color channels, we obtain

$$\mathbf{H} = \begin{bmatrix} \mathbf{Z}_{xx}^R & \mathbf{Z}_{xy}^R \\ \mathbf{Z}_{yx}^R & \mathbf{Z}_{yy}^R \\ \mathbf{Z}_{xx}^G & \mathbf{Z}_{xy}^G \\ \mathbf{Z}_{yx}^G & \mathbf{Z}_{yy}^G \\ \mathbf{Z}_{xx}^B & \mathbf{Z}_{xy}^B \\ \mathbf{Z}_{yx}^B & \mathbf{Z}_{yy}^B \end{bmatrix}, \quad \mathbf{z}_t = \begin{bmatrix} \mathbf{z}_{xt}^R \\ \mathbf{z}_{yt}^R \\ \mathbf{z}_{xt}^G \\ \mathbf{z}_{yt}^G \\ \mathbf{z}_{xt}^B \\ \mathbf{z}_{yt}^B \end{bmatrix} \tag{S4}$$

in which we define the sub-matrices of $\mathbf{H}$ and $\mathbf{z}_t$ similarly to before.

Revaud *et al.* [4] use a weighted combination of two data terms $d(\mathbf{u})$ based on (S2) and (S4). This combination can be understood as forming new $\mathbf{H}$ and $\mathbf{z}_t$ by stacking the ones in (S2) and (S4). When the two data terms are combined using equal weights, the inverse covariance matrix $\mathbf{H}^*\mathbf{H}$ becomes

$$\mathbf{Z} = \begin{bmatrix} \sum \mathbf{Z}_{*x}\mathbf{Z}_{*x} & \sum \mathbf{Z}_{*x}\mathbf{Z}_{*y} \\ \sum \mathbf{Z}_{*x}\mathbf{Z}_{*y} & \sum \mathbf{Z}_{*y}\mathbf{Z}_{*y} \end{bmatrix}, \tag{S5}$$

and the transformed signal is

$$\mathbf{H}^\dagger \mathbf{z} = \begin{bmatrix} \sum \mathbf{Z}_{*x}\mathbf{z}_{*t} \\ \sum \mathbf{Z}_{*y}\mathbf{z}_{*t} \end{bmatrix}, \tag{S6}$$

cf. (27) in our main paper. In (S5)–(S6), the summations are over the three color channels for each of the 0th, and the 1st partial derivatives of the image. Figure 1 visualizes our flow estimates.

1.3. Image Deblurring

Figure 2 provides crops of the deblurred images from the the Kodak dataset [2], produced by different algorithms. We optimize the algorithm parameters for the different methods (Wiener, $L2$, and TV) via grid search. The Wiener filter uses a uniform image power spectrum model. Note the use of the bilateral filter is not optimal for de-noising as pointed out by Buades *et al.* [5], who demonstrate the advantages of patch-based filtering (nonlocal means denoising) over pixel-based filtering (bilateral filter). Our deblurring results are based on the bilateral filter, but one is free to use the non-local means filter (or any other filter) for the de-noising operator $\mathbf{A}$.

Figure 1: Optical flow (top rows) and the corresponding flow error (bottom rows) produced using the geodesic and the bilateral variants of our method. Whiter pixels correspond to smaller flow vectors.

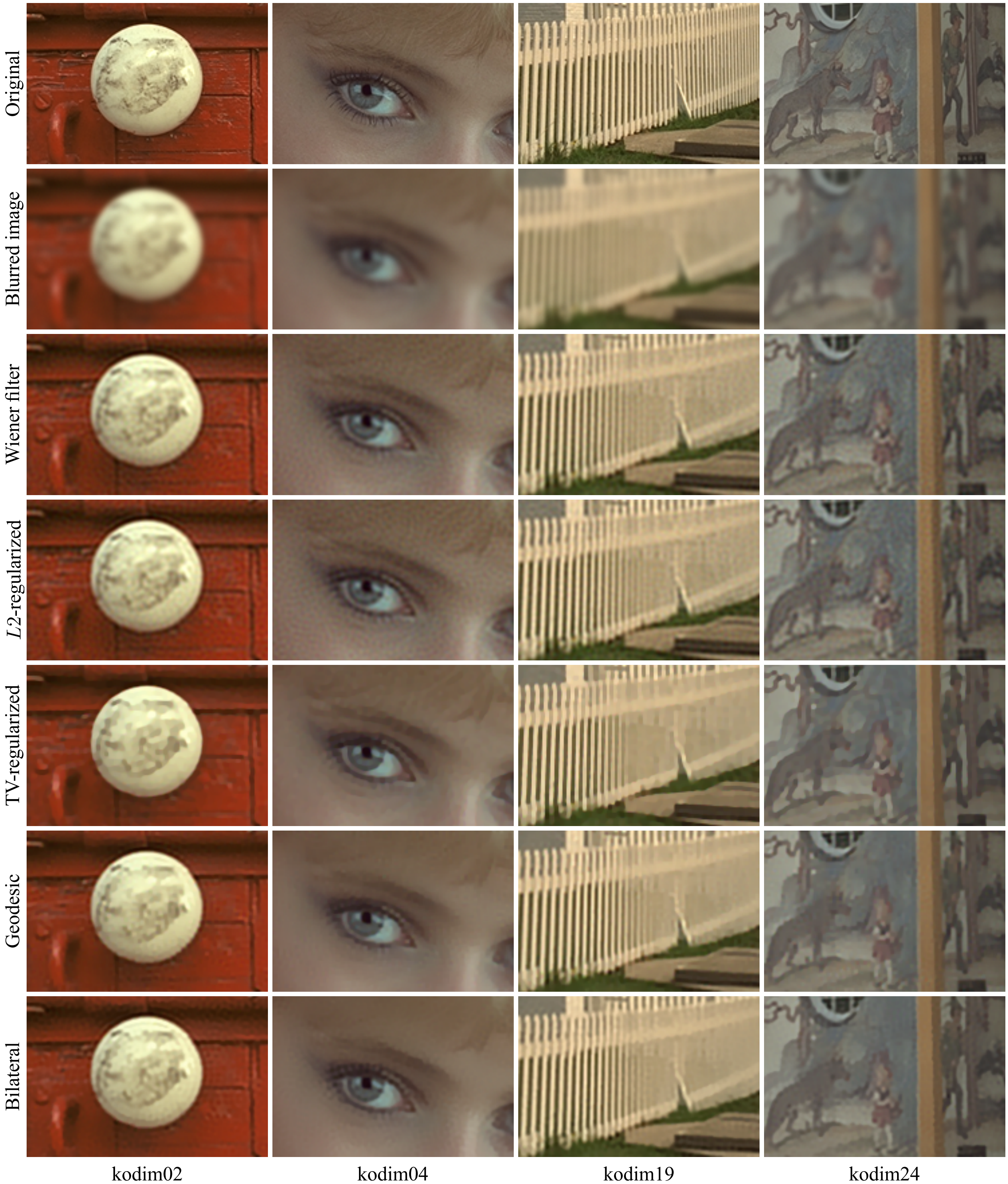

Figure 2: Crops of images from the Kodak dataset when the B-spline blur kernel ($n = 8$) is used. Our method exhibits less ringing compared to the Wiener filter and the $L2$-regularization methods, and has less staircasing artifacts than the $L1$ (TV) method.

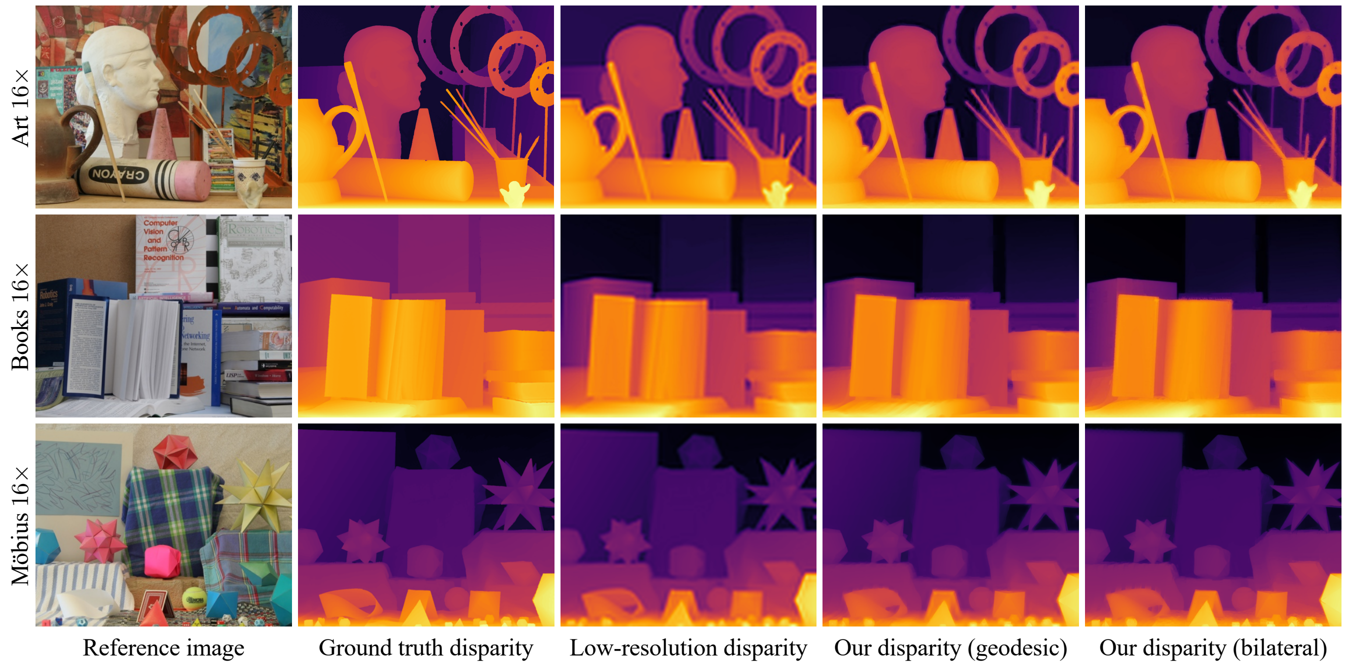

Figure 4: The $16\times$ super-resolution disparity maps produced using the geodesic and the bilateral variants of our method for the 1088×1376 scenes Art, Books, and Möbius used in [1]. Best viewed online by zooming in.

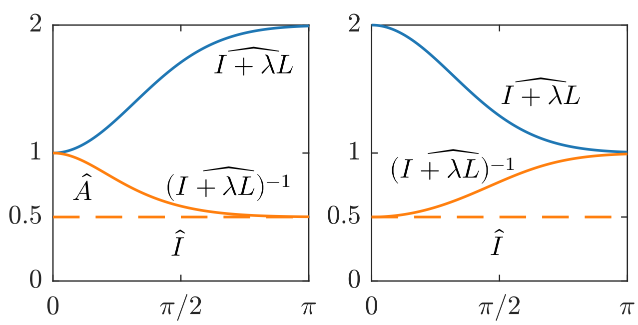

Figure 3. The frequency response $(\widehat{C+\lambda L})^{-1}$ can be expressed as a sum of low-pass response $\widehat{A}$ and an all-pass one $\widehat{I}$ only when the response $(\widehat{C+\lambda L})^{-1}$ is low-pass-like (left). Shown for $\lambda = 1$.

2. Possible Limitations

In Section 4 of our paper, we discussed that (14a) is valid only when (14a) matrix $(\mathbf{C} + \lambda\mathbf{L})^{-1}$ has a low-pass spectral response. We show this in Figure 4 (left) for the case where $\lambda = 1$ and $\mathbf{C} = \mathbf{I}$. Since $\mathbf{C} + \lambda\mathbf{L}$ is Sinkhorn-normalized, it has a high-pass spectral response $\widehat{I+\lambda L}$, ranging from 1 to 2. As a consequence, the inverse filter response $(\widehat{I+\lambda L})^{-1}$ ranges from 1 down to 0.5. We can approximate such a filter response as a sum of low-pass and all-pass responses. In our context, an approximation of $\mathbf{u}^{\mathrm{opt}} = (\mathbf{C} + \lambda\mathbf{L})^{-1}\mathbf{Cz}$ can be obtained using a convex combination of $\mathbf{Cz}$ and a low-pass-filtered version $\mathbf{ACz}$ of it. On the other hand, if $\widehat{I+\lambda L}$ is a low-pass response. In this case, the inverse response (shown in Figure 4, right) is high-pass, and the solution $\mathbf{u}^{\mathrm{opt}}$ cannot be approximated as a convex combination of $\mathbf{Cz}$ and a low-pass-filtered version of it. In practice, we can still use (14b) to solve the transformed problem.

References

[1] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S. Brown, and In So Kweon. High quality depth map upsampling for 3D-TOF cameras. In *ICCV*, 2011.

[2] Nils Papenberg, Andrés Bruhn, Thomas Brox, Stephan Didas, and Joachim Weickert. Highly accurate optic flow computation with theoretically justified warping. *Int. J. Comput. Vis.*, 67(2):141–158, 2006.

[3] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artif. Intell.*, 17(1):185–203, 1981.

[4] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. EpicFlow: Edge-preserving interpolation of correspondences for optical flow. In *CVPR*, 2015.

[5] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A non-local algorithm for image denoising. In *CVPR*, 2005.