Vision, Modeling, and Visualization (2019), pp. 1–9
H.-J. Schulz, M. Teschner, and M. Wimmer (Eds.)

Consistent Filtering of Videos and Dense Light-Fields Without Optic-Flow *

Sumit Shekhar[1], Amir Semmo[1,2], Matthias Trapp[1], Okan Tarhan Tursun[3,4], Sebastian Pasewaldt[1,2], Karol Myszkowski[3], and Jürgen Döllner[1]

[1]Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany
[2]Digital Masterpieces GmbH, Germany
[3]Max-Planck Institute for Informatics, Germany
[4]Università della Svizzera italiana Lugano, Switzerland

Abstract

A convenient post-production video processing approach is to apply image filters on a per-frame basis. This allows the flexibility of extending image filters—originally designed for still images—to videos. However, per-image filtering may lead to temporal inconsistencies perceived as unpleasant flickering artifacts, which is also the case for dense light-fields due to angular inconsistencies. In this work, we present a method for consistent filtering of videos and dense light-fields that addresses these problems. Our assumption is that inconsistencies—due to per-image filtering—are represented as noise across the image sequence. We thus perform denoising across the filtered image sequence and combine per-image filtered results with their denoised versions. At this, we use saliency based optimization weights to produce a consistent output while preserving the details simultaneously. To control the degree-of-consistency in the final output, we implemented our approach in an interactive real-time processing framework. Unlike state-of-the-art inconsistency removal techniques, our approach does not rely on optic-flow for enforcing coherence. Comparisons and a qualitative evaluation indicate that our method provides better results over state-of-the-art approaches for certain types of filters and applications.

CCS Concepts

• ***Computing methodologies*** ,..., ***Image processing;*** *Computational photography;*

1. Introduction

Due to rapid advancements in the field of visual computing in the past few decades, a plethora of image-processing techniques have been developed that deal with manifold applications, such as tone-mapping, contrast enhancement, color constancy, color grading, and style transfer. However, extending such techniques for *video* is not a trivial task. The difficulty arises due to an extra *temporal* dimension in the input data. The *local* image-processing methods—generally implemented in the form of 2D kernels—often lead to spatial and temporal inconsistencies when extended to 3D kernels [BTS*15, LHW*18]. These inconsistencies may appear due to two main reasons: 1) different distributions of image features between adjacent video frames, and 2) considering only a small temporal window when filtering. The extension of *global* image processing methods would require processing a video as a whole, which might not be feasible in real-time using conventional hardware or in streaming scenarios.

One naive, yet generic way of extending image-based filtering techniques for video is to apply them individually on a per-frame basis. However, this approach may lead to temporal incoherences (Fig. 1(b)). A more sophisticated approach is to implement application-specific techniques for video, e.g., tone-mapping [ASC*14], color grading [BSPP13, YCC16], color constancy [FL11, BT17], intrinsic decomposition [BST*14, MZRT16], where application-specific constraints are employed. At this, however, one needs to be familiar with the particular filtering technique, and specific approaches might not be applicable to other filters.

A more generalized approach is followed in previous methods that are agnostic to the type of filtering [Par08, LWA*12, BTS*15, YCC17, LHW*18]. These works are based on the idea of performing per-frame filtering and applying temporal consistency as a constraint during processing or as a post-processing step. Most of these techniques implicitly require optic-flow. For instance, Bonneel *et al*. [BTS*15] use flow-based image warping in a gradient-domain-based optimization to enforce consistency between neighboring views, and Lai *et al*. [LHW*18] use optic-flow to train a neural network by minimizing short-term and long-term temporal loss for enforcing consistency—which also applies to *neural style transfer* (e.g., [GEB16, JAFF16]) for video.

submitted to *Vision, Modeling, and Visualization (2019)*

However, optic-flow computations may be expensive and/or potentially inaccurate especially in case of disocclusion [FBK15]. In this work, we present a consistent filtering technique for image sequences that does not rely on optic-flow and still can attain temporal consistency in a generalized way (Fig. 1). In particular, we show that a careful combination of low-frequency content from the temporally denoised output and high-frequency content from the per-frame result can significantly reduce temporal flickering. We use saliency-based weights for such an adaptive combination, i.e., to identify and preserve visually important details.

Unlike most of the previous methods, our algorithm is well suited for image-abstraction applications e.g., *neural style transfer*. Moreover, our method is applicable to filter dense light-fields, which gained major attention in the past decade [WMJ*17, OEE*18] with the advent of Virtual Reality (VR). Manifold image processing methods [WMJ*17] have been extended to dense light-fields for applications such as denoising [AS17, WG14, MV12], intrinsic decomposition [AG16, GEZ*17, BSM*18], and depth estimation [JPC*15, JHG*17, SDRJ18]. Our video-based solution is applicable to a wide variety of image filters and can be easily extended to dense light-fields. To summarize, this work makes the following contributions:

1. A method that makes per-image filtered image sequences consistent by denoising image slices across the sequence *without using optic-flow*.
2. An interactive real-time processing framework that enables direct control of the amount of temporal or angular consistency, based on image saliency, thus producing user-defined outputs.
3. Applications demonstrate the versatile usage of our method to a wide-range of image filters for video and dense light-fields such as color grading, color constancy, dehazing, colorization, and neural style transfer.

2. Related Work

2.1. Task-Specific Consistent Video Filtering

Many application-specific techniques have been extended to achieve temporal consistency based on the type of image filter. For instance, Aydin *et al.* [ASC*14] propose to use edge-aware spatio-temporal filtering of High Dynamic Range (HDR) videos to obtain *base* and *detail* layers and perform coherent video tone-mapping. Temporal coherence is a particular challenge for video tone-mapping as surveyed by Eilertsen *et al.* [EMU17]. For the application of color grading, Bonneel *et al.* [BSPP13] employ an approximate curvature-flow technique to enforce temporal consistency in a post-processing step. In the context of color constancy, Farbman *et al.* [FL11] use the tonal settings of few *anchor frames* to process in-between frames to ensure consistency. In case of video stylization optic-flow is typically used for automated coherent parameterization, e.g., in bi-directional texture advection of watercolor stylizations [BNTS07], to compute *object flow*—robust against inaccurate optic-flow—for generalized video stylization [LXT17], and in machine learning for coherent style transfer [RDB18]. For intrinsic decomposition Meka *et al.* [MZRT16] use a global spatio-temporal reflectance consistency prior ensuring temporal consistency. The above application-specific examples show the variety of techniques used to overcome the common underlying problem of temporal inconsistency. Most of them utilize optic-flow to enforce temporal coherence. Unlike the above approaches, we develop a generic algorithm that is filter or task agnostic. Moreover, our method does not rely on optic-flow.

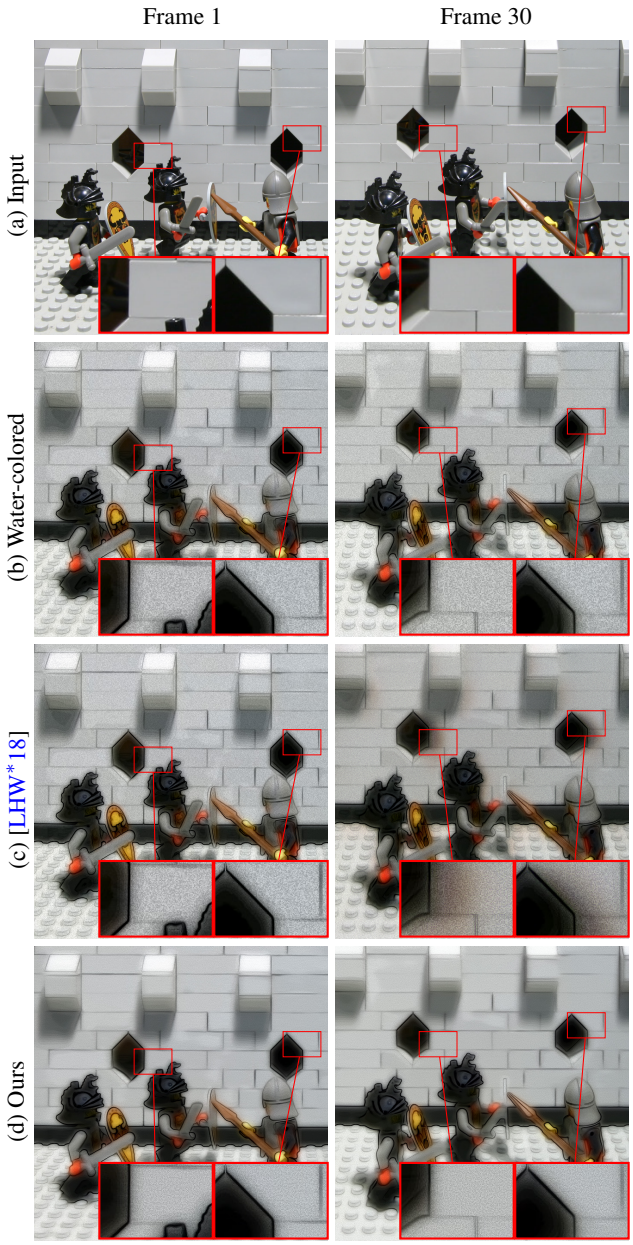


Figure 1: Comparison of angular (in-)consistency for (a) *Lego* light-field (taken from [VA08]) processed with (b) per-frame water-color stylization using the method of Bousseau *et al.* [BNTS07]. As can be observed in this example, (c) the output produced with the technique of Lai *et al.* [LHW*18] introduce visible artifacts as compared to (d) our approach that provides more consistent result.

2.2. Task-Agnostic Consistent Video Filtering

Apart from application-specific approaches, generic methods have been proposed that solve the problem of temporal inconsistency for various filters. Paris [Par08] extends image-based isotropic dif-

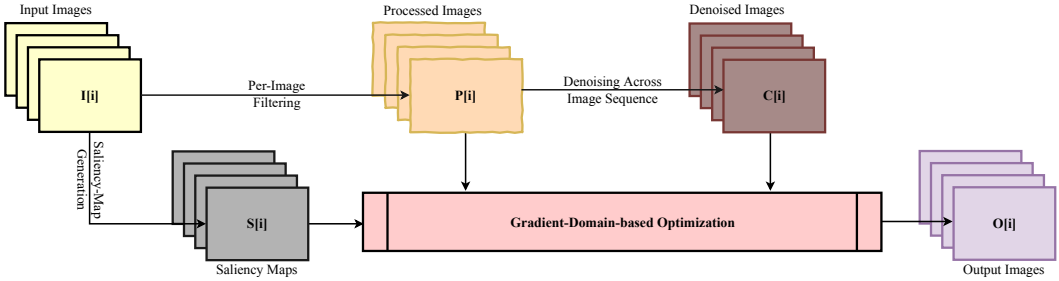


Figure 2: Flowchart of our system for consistent filtering of an image sequence as described in Sec. 3.

fusion and Gaussian convolution for video streams with an application towards bilateral filtering, anisotropic diffusion, and mean-shift segmentation. Lang *et al.* [LWA*12] create motion paths using dense optic-flow, which are then filtered after undergoing a 1D domain transform. Dong *et al.* [DBZY15] divide individual frames of a video into multiple regions and perform a region-based spatio-temporal optimization. Bonneel *et al.* [BTS*15] combine the high-frequency gradients from the per-frame processed output and the low-frequency content from the warped version of the previous frame using a gradient-domain based optimization scheme. Yao *et al.* [YCC17] use key frames to avoid the inconsistency problem that occur due to occlusion. Finally, Lai *et al.* [LHW*18] use a machine-learning technique and introduce short-term and long-term temporal losses as well as a perceptual loss to balance temporal coherence between frames and perceptual similarity with the individually processed frames. Our method also belongs to this category of generic approaches to attain the goal of temporal consistency, but—unlike previous methods—does not require optic-flow.

2.3. Light-Field Filtering

Many generic methods have been recently proposed to propagate per-view edits consistently across dense light-fields [JMG11, AZJ*15, FG17]. Jarabo *et al.* [JMG11] downsample the light-field data based on an affinity function. The edits are propagated in the downsampled domain. Ao *et al.* [AZJ*15] build upon the work of Jarabo *et al.* and perform an improved downsampling and upsampling on reparameterized light-fields to explicitly enforce consistency between views. Frigo *et al.* [FG17] perform diffusion in the Epipolar Plane Images (EPIs) for an angularly-coherent light-field editing. In a follow-up work, Bonneel *et al.* [BTS*17] extend their previous work [BTS*15] on single-camera videos to multi-camera array videos, which is also applicable to light-fields.

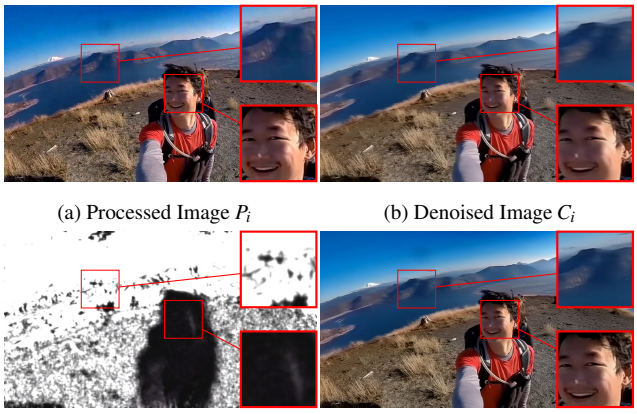
These techniques are examples on how to approach the common problem of angular inconsistency. However, for light-fields we aim to preserve angular consistency analogous to temporal consistency in videos. Our approach for the removal of temporal inconsistencies can be extended to achieve angular consistency for light-field filtering with only minor modifications. Moreover, in case of light-fields, our denoising step corresponds to EPI denoising and such EPI manipulation is an integral aspect of various light-field processing methods [WMJ*17].

3. Method

For an input image sequence $\{I_i \mid i = 1 \dots N\}$, its per-image processed version $\{P_i \mid i = 1 \dots N\}$, and per-image saliency map $\{S_i \mid i = 1 \dots N\}$, we seek to find a consistent output $\{O_i \mid i = 1 \dots N\}$. Our method is agnostic to the filter f applied on each image. As an intermediate step, P_i is denoised across the image sequence (Secs. 3.1 and 3.2) to obtain $\{C_i \mid i = 1 \dots N\}$. We then solve a gradient-domain optimization scheme in the image domain Ω (Fig. 2),

$$E(O_i) = \int_\Omega \left(\underbrace{||\nabla O_i - \nabla P_i||^2}_{\text{data}} + \underbrace{w_s||O_i - C_i||^2}_{\text{smoothness}} \right) d\Omega \quad (1)$$

The *data* term in this optimization approach enforces similarity with the per-image processed result P_i in the gradient-domain. Thus, only the high-frequency details are taken from P_i. The low-frequency consistent content is taken from the denoised image C_i. The influence of *smoothness* term is controlled by per-pixel saliency weights w_s (Fig. 3).



(a) Processed Image P_i (b) Denoised Image C_i

(c) Perceptually-salient Weight w_s (d) Output O_i

Figure 3: Adaptive combination of a per-frame processed image and its denoised version using perceptually-salient weights. Note how the perceptually salient face details in foreground are preserved in the output while the background is temporally smoothed.

3.1. Temporal Denoising

Our assumption is that the temporal inconsistencies in a video are represented as temporal noise across a given scanline. We arrange the video frames of P_i to form an image sequence where—apart from the spatial dimension—the third dimension represents time. The image sequence is horizontally sliced across a given scanline to obtain a respective TSI (Fig. 4). The temporal inconsistencies in

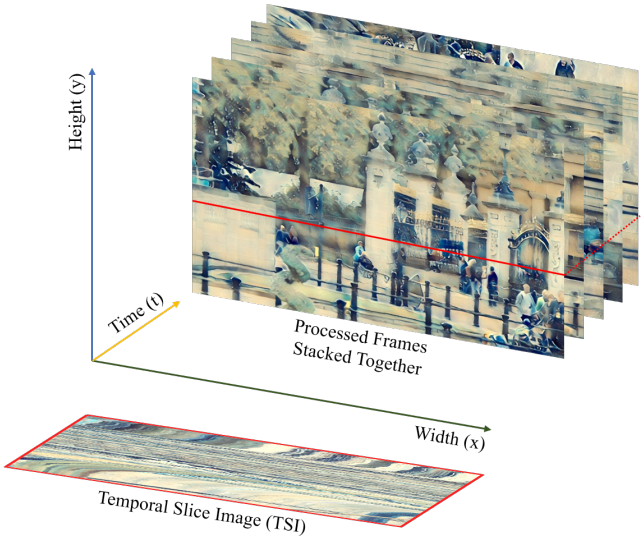


Figure 4: Exemplary processed image sequence and its corresponding Temporal-Slice Image (TSI).

the video can be seen as noise in the TSI (Fig. 5b). A straightforward approach to remove inconsistencies is to perform denoising in the TSI domain (Fig. 5). A side-effect of the denoising step is the introduction of motion blur along the horizontal direction (Fig. 3b). It is also possible to slice the image sequence vertically, thereby causing blur in the vertical direction. However, the direction of slicing does not affect the final output noticeably (Fig. 6). Our approach of denoising image slices is inspired from the work of Khazdan *et al*. [KBK*13], where the authors use a similar technique for the denoising of electron microscopy image stacks.

In order to denoise a temporal slice, a method of choice should be the one that reduces temporal inconsistencies without introducing motion-blur in the image sequence. We experimented with four image denoising methods for this purpose: naive Gaussian smoothing, Bilateral filtering [TM98], BM3D [DFKE07], and FFDNet [ZZZ17]. In comparison to others the learning-based denoising of FFDNet can handle spatially variant noise, wide range of noise levels and is also fast. It is based on an end-to-end trainable deep CNN that incorporates residual learning. Moreover, we empirically identified FFDNet to be the best choice for our use case w.r.t the above mentioned criteria (Fig. 5).

3.2. Angular Denoising

In case of dense light-fields, the third dimension in the stacked image sequence represents the angular dimension. The sequence of processed sub-aperture views are traversed in horizontal and vertical directions from the top-left to bottom-right to obtain *horizontally* and *vertically* traversed image-sequences respectively (Fig. 7). Each of these image sequences is sliced along a given scanline to obtain an Angular-Slice Image (ASI) comprising of multiple EPIs (Fig. 8). The sliced images—representing the EPI domain—are denoised for removing angular inconsistencies. The denoised horizontal and vertical traversed image sequences are averaged to obtain the final angularly-denoised light-field. We employ the same denoising algorithm as in case of temporal denoising.

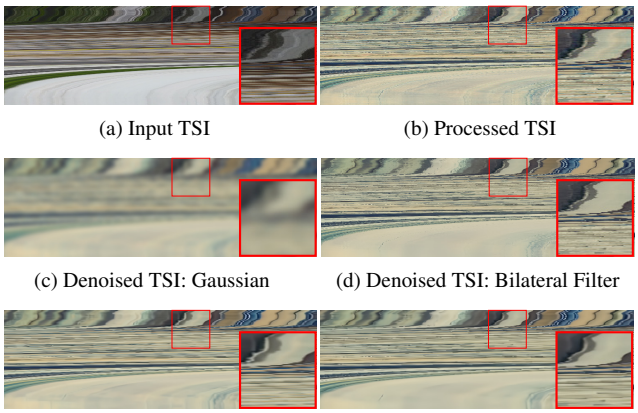


Figure 5: Example of a TSI for the correspondent (a) input, (b) processed, and denoised videos. Compared denoised versions: (c) Gaussian, (d) Bilateral Filter, (e) BM3D, and (f) FFDNet.

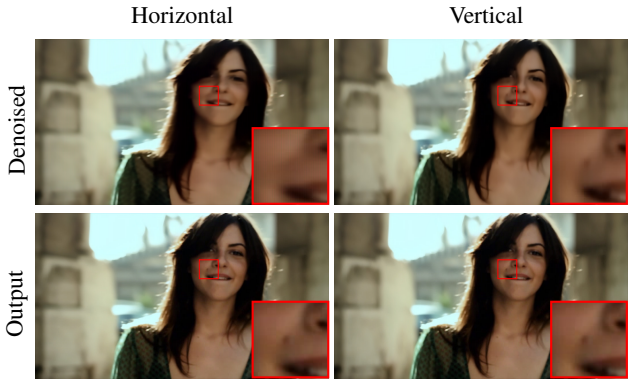


Figure 6: Comparing the outputs after denoising (top row) and outputs after consistent filtering (bottom row) with respect to computing horizontal and vertical TSIs.

3.3. Saliency Weight

The minimization of the energy function in Eqn. (1) aims to achieve two main goals: (a) perceptual-similarity with the per-image processed result and (b) reduced inconsistencies. The consistent image C_i is smoothed due to denoising; however, such smoothing also blurs image details. To enforce consistency and also preserve important details, we make use of a per-pixel perceptually salient weight w_s. The idea is to allow for more smoothing in those regions that are either not salient $(1-S_i)$ or where the difference in intensities of P_i and C_i is not noticeable $(1-D_i)$ (Eqns. (2) to (3)). The scaling and offset parameters $\beta \in [0.1, 10.0]$ and $\varepsilon \in [0.02, 1.0]$ facilitate tuning the weight, respectively:

$$w_s = \beta[(1-S_i)(1-D_i)+\varepsilon] \tag{2}$$

The definition of the binary just-noticeable-difference function D_i uses a *diff* value threshold. The threshold parameter $\mu \in [0.01, 10.0]$ provides further tuning control. The *diff* function (Eqn. (4)) is based on the definition of Weber contrast and uses image intensity

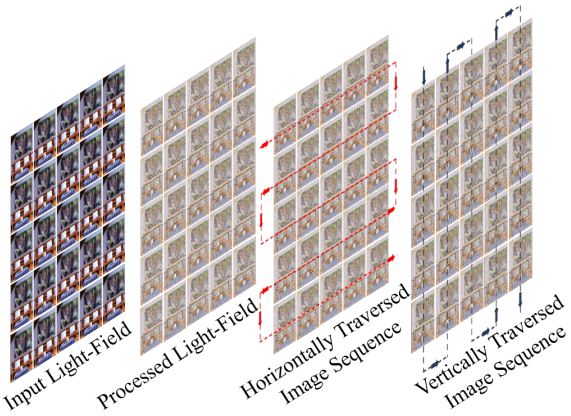


Figure 7: Schematic overview of how the image sequences are horizontally and vertically traversed.

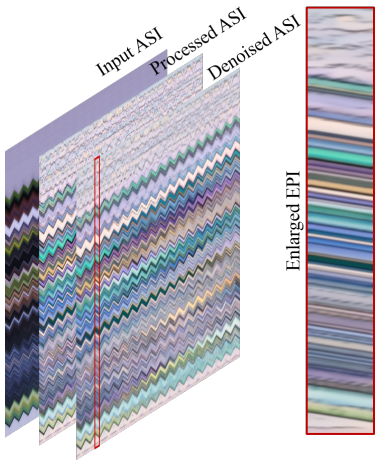


Figure 8: Exemplary overview of an input, processed and denoised ASI. The inset shows an enlarged EPI.

as a measure [Web96]. In this respect, we observe that the image intensity measure (Eqn. (5)) empirically performs better than the luminance for the purpose of consistency [NB17].

$$D_i = \begin{cases} 1, & \text{if } diff \geq \mu \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$diff = \frac{|In(P_i) - In(C_i)|}{In(P_i)} \tag{4}$$

$$In(\,I_i\,) = \sqrt{r^2 + g^2 + b^2} \tag{5}$$

In order to compute saliency maps, we experimented with (1) the image-based method of Liu *et al*. [LHY18] and (2) the video-based method of Wang *et al*. [WSS18]. Here, we observe that the spatial resolution of saliency maps are better with (1) while the temporal consistency is better with (2), we thus favor the technique of Wang *et al*. for our purposes. The resultant weight is smoothed with a Gaussian filter ($\sigma \in [0.1, 5.0]$) to improve its spatial consistency. By tuning the parameters we make sure that saliency weights vary smoothly between frames.

Algorithm 1 Consistent Filtering of a Video-Sequence

for $i \leftarrow 1$ to N **do** ▷ N number of images
  $P_i \leftarrow f(I_i)$ ▷ Per-image filtering
for $k \leftarrow 1$ to H **do** ▷ H is height (in pixels) of each image
  $TSI_k \leftarrow Slice(\{P_i \mid i = 1 \ldots N\})$ ▷ Slice across P_i sequence
  $Denoise(TSI_k)$
for $i \leftarrow 1$ to N **do**
  $C_i \leftarrow MergeSlices(\{TSI_k \mid k = 1 \ldots H\})$
  $S_i \leftarrow ComputeSaliency(I_i)$
  $w_s \leftarrow ComputeSaliencyWeights(S_i, \beta, \varepsilon)$
  $O_i \leftarrow SolveOptimization(P_i, C_i, w_s)$

3.4. Optimization Solver

The output O_i, which minimizes the energy $E(O_i)$ in Eqn. (1), must satisfy Eqn. (6) as per the Euler-Lagrange formulation [Wei74]:

$$w_s \cdot O_i - \Delta O_i = w_s \cdot C_i - \Delta P_i \tag{6}$$

For solving the system of linear equations represented by Eqn. (6), we use the iterative scheme of *Stochastic Gradient Descent (SGD)* with *momentum* [Qia99]. By choosing an iterative solver, we overcome the limitation of storing a large matrix in memory and calculating its inverse. Moreover, with an iterative scheme we can stop the solver once we have achieved a solution without noticeable inconsistencies. At this, our interactive interface allows users to control the degree of convergence by providing the number of iterations. In practice, with a fast convergence rate of *SGD with momentum*, 20 - 30 iterations are sufficient for a consistent output. Our technique for consistent filtering, summarized in Algo. 1, does not require optic-flow. However, we can extend our optimization (Eqn. (1)) to include potentially accurate flow information (see supplementary material).

4. Results

Our approach is independent of the underlying image filtering applied on the video frames or light-field sub-aperture views, and is suitable for a wide range of applications (Fig. 10 and Fig. 12).

Neural Style Transfer. We apply the feed-forward neural style transfer of Johnson *et al*. [JAFF16] per video frame. Using our consistent filtering approach, high-frequency temporal flickering can be reduced in the stylized output. Recent works argue that many filtering approaches have become too successful at coherent stylization, as the outputs loose the "visual richness comparable to real artwork" [FLJ*14] or have "the uncanny and unappealing effect of a 3D world covered in paint" [DBH19]. In this respect, the proposed interactive framework and saliency maps can help to locally control the amount of temporal or angular consistency, and thus preserve detail of the transferred style.

Image enhancement. In order to enhance individual images, we use the low-light image enhancement technique by Chen *et al*. [WWYL18]. The per-image operation introduces high-frequency flickering like film-grain noise. Our method provides inherent denoising and is able to provide a consistent output.

Colorization. We use the image colorization algorithm of Zhang *et al*. [ZIE16] to colorize individual frames of a video.

Table 1: Statistics of the Likert scale score for evaluated techniques

	[BTS*15]	[LHW*18]	Ours
Mean (μ)	2.49	3.06	**3.56**
Std. Error of Mean (σ_μ)	0.10	0.08	**0.06**

The temporal flickering is caused due to color variations between frames as well as color bleeding within scene objects. Our methods is able to significantly reduce these artifacts.

Color Grading. Applying the first part of the color grading algorithm proposed by Bonneel *et al*. [BSPP13] to videos results in obvious temporal inconsistencies. These can be noticeably removed using our method.

HDR Toning. We apply tone-mapping using the method of Paris *et al*. [PHK11] on a per-frame basis. The toning technique—based on subband decomposition—causes flickering in consecutive frames due to different high and low frequency luminance details. Our algorithm is able to rectify these luminance variations between separately tone-mapped images.

4.1. Comparative Evaluation

We compare our algorithm with the previous methods of Bonneel *et al*. [BTS*15] and Lai *et al*. [LHW*18] for the above mentioned applications (Fig. 10 and Fig. 12). In case of videos, we use the test dataset provided by Lai *et al*. for relative comparison and for light-fields we generate the corresponding results. We observe that the method of Bonneel *et al*. is not suitable for applications where new image edges are generated as part of the filtering process, e.g., stylization and neural style transfer. Moreover, since their method is based on the accuracy of the optic-flow, they suffer from artifacts when occlusion occurs in large spatial regions (Fig. 12(b)). Since our approach does not require optic-flow, it is robust to problems due to occlusion and can also handle creation of new edges. The approach of Lai *et al*. addresses the problem of occlusions by introducing a long-term temporal loss. However, such long-term loss also propagates the inconsistencies from temporally or angularly distant frames. We observed such inconsistency propagation in the form of subtle luminance or color variations (Fig. 1 and Fig. 10). In comparison, our approach is based on denoising of TSI or ASI images using a local-denoising method that does not affect regions that are spatially distant in the TSI or ASI domain.

4.2. User Study

We conducted a user study to qualitatively evaluate the output of our approach. We ask the participants to watch the consistent outputs and rate them on a Likert scale.

Setup. For each scenario we show a participant five videos, two on the top row and three on the bottom. On the top row, we have the original video and its per-frame processed version. We make the per-frame processed video consistent using our, Boneel *et al*. [BTS*15], and Lai *et al*. [LHW*18] methods and place them in the bottom row. The order of videos in the bottom row is randomized for each sample. At first the videos in the top row are played while those in the bottom row are stopped. After the user has seen the top row videos, bottom row videos are played. The videos are played continuously in a playback loop. For each case,

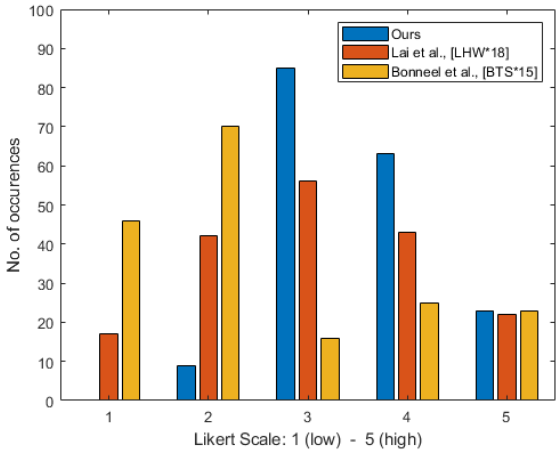


Figure 9: Likert scale score of ours and previous methods as per the user study (Sec. 4.2).

participants were asked to rate the overall visual quality of outputs on a Likert scale from 1 (low) to 5 (high) based on two criteria: (1) the consistency of the output and (2) its resemblance with the per-frame result. A total of 18 people (4 female, 13 male, 1 no answer) within an age group of 20 - 40 participated in the above study and each looked at 10 (6 video and 4 light-field) filtering examples. The distance between the screen and the observer was fixed to 1 m for all participants. The group of participants included users with and without prior knowledge of image and video processing.

Analysis. In comparison to others, our method was able to improve over the per-frame result for most of the cases (Fig. 9). In case of light-fields, we perform significantly better than the previous methods (see supplementary material). We compute mean and standard error of mean of Likert scale scores (Table 1) and perform "Two-Sample t-Test for Equal Means" for validation. We observe a significant difference between the average scores of our method vs Lai *et al*. and Boneel *et al*. ($p < 0.005$, t-test) respectively.

4.3. Performance

All our experiments were performed on a PC using Microsoft Windows 7 as operating system, with a 3.5 GHz CPU, 16 GB of RAM, and a Nvidia GTX 1050 Ti graphics card with 4 GB VRAM. The processed images are denoised to obtain C_i and the saliency maps S_i are computed in a pre-processing step. The denoising of image slices is implemented in Python using the PyTorch [TDV19] reference implementation of the FFDNet [ZZZ17]. For a video sequence of 219 frames, each with a spatial resolution of 1024×576 pixels, computing C_i takes approx. 50 seconds for all frames. The dynamic video saliency map is computed using the original implementation by Wang *et al*. [WSS18], which takes approx. 135 seconds for all frames. Our interactive system is able to perform steps 6 to 10 of Algo. 1 in real-time for each frame. It is implemented with C++ and CUDA (v10.0) and takes 30 to 35 milliseconds per-frame to perform 30 iterations of *SGD with momentum* to solve Eqn. (6).

5. Discussion

Our findings suggest that a careful combination of per-image processed results and their temporal/angular denoised versions can be

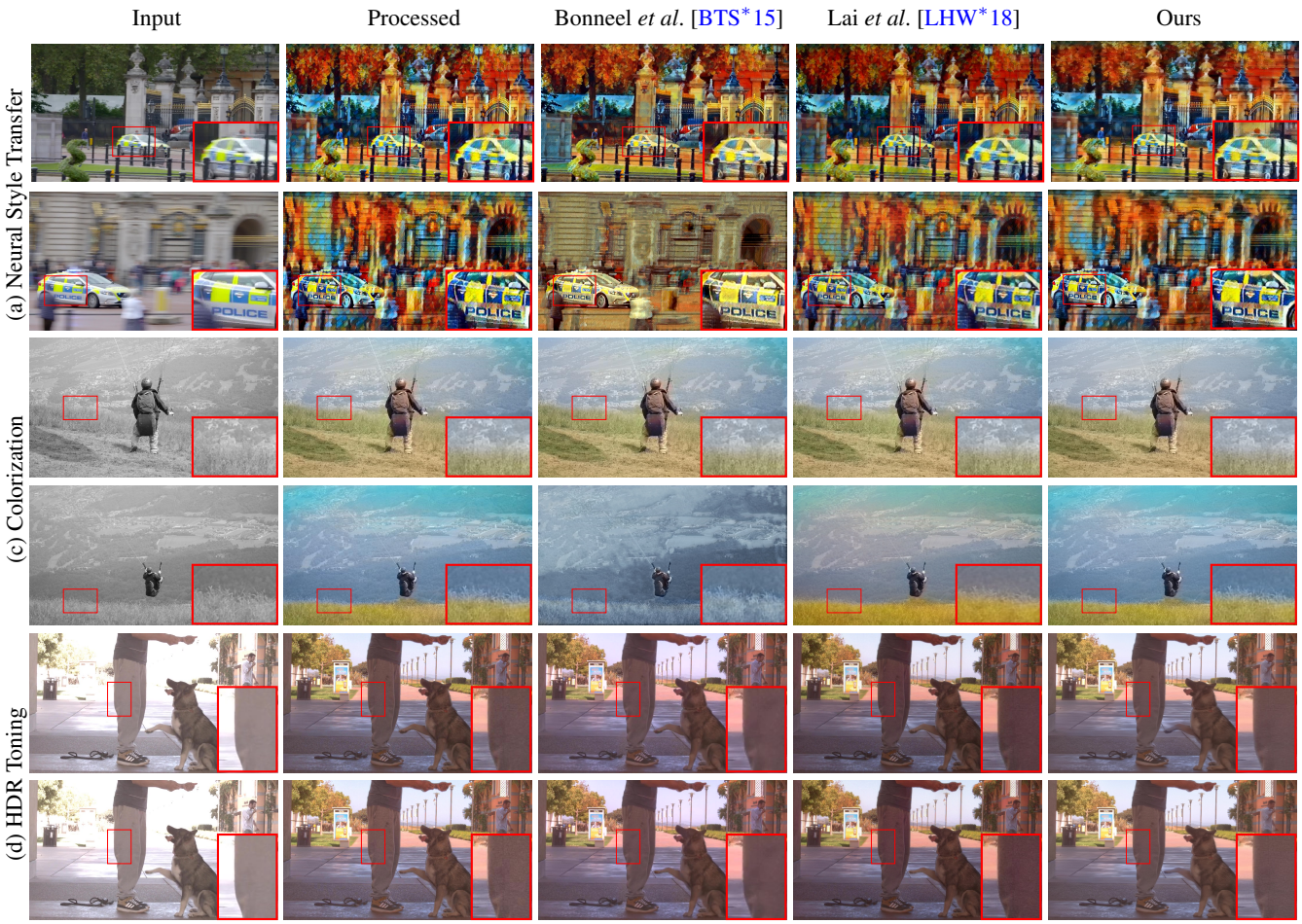


Figure 10: Comparison of our video consistency filtering technique with previous methods by applying per-frame (a) Neural Style transfer by Johnson *et al*. [JAFF16] (b) Color constancy by Gijsenij *et al*. [GGvdW12] (c) Image-colorization by Zhang *et al*. [ZIE16] (d) HDR Toning by Paris *et al*. [PHK11]. Please refer to supplementary video for better visualization. Input videos are taken from the work of Bonneel *et al*. [BTS*15] and the *DAVIS* dataset [PPTM*16].

used to generate perceptually consistent outputs. It implies that selective denoising of a processed image sequence is effective in removing noticeable inconsistencies. In comparison, previous methods mainly relied on optic-flow based image warping of consecutive frames for enforcing consistency. The image-based warping technique might not be effective for cases where optic-flow computation is challenging.

Our algorithm performs consistent filtering based on denoising of TSI or ASI images. The above denoising step requires the complete sequence as an input and can only be applied as a post-processing step. Thus, our approach is not suitable for video streaming applications. We use carefully designed optimization weights to strike a balance between preserving details and enforcing consistency. However, we believe that this trade-off can be further improved by performing a thorough analysis of the spatio-temporal/spatio-angular contrast sensitivity [DAC10]. We believe that a saliency-map which has a higher spatial resolution and better temporal consistency can further enhance our results.

We perform horizontal slicing of image and also evaluate vertical slicing in the denoising step. As part of future work, we would analyze stochastic sampling of both the horizontal and vertical neighborhood to avoid any potential residual bias. In case of large or arbitrary movement of objects the denoising approach cannot avoid introducing noticeable motion blur Fig. 11. However, even in such cases we perform relatively better than previous methods (see supplementary material).

6. Conclusions

In this work, we propose an algorithm to reduce incoherencies in per-frame filtering of image sequences without relying on optic-flow. At this, denoising is performed across image sequences in a pre-processing stage and a least-squares energy minimization is solved in real-time. By carefully designing optimization weights, the algorithm is able to preserve visual details and maintain coherence in videos and dense light-fields. Our results for image and video processing techniques demonstrate that our approach is filter-agnostic and indicate improved output quality over state-of-the-art methods for certain types of filters and popular applications. As part

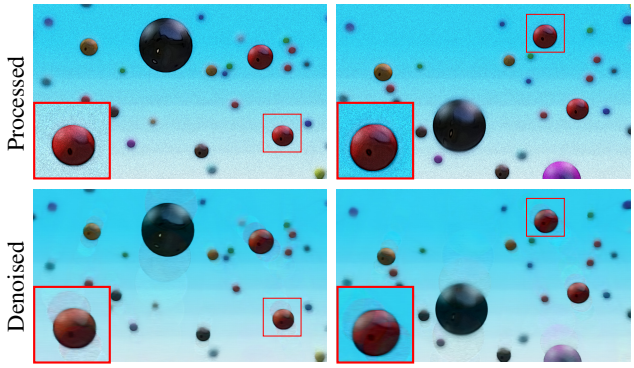


Figure 11: The per-frame processed and denoising output where objects are moving in arbitrary trajectory. We observe motion artifacts similar to ghosting along the trajectory of spheres.

of future work, we plan to make our approach causal and applicable to streams of image sequences.

Acknowledgements

We thank Florian Wagner for his help with generating results. We thank Max Reimann for valuable discussion. We thank the reviewers for their insightful comments. The project was supported by the Fraunhofer and Max Planck cooperation program within the German pact for research and innovation (PFI), the Research School on "Service-Oriented Systems Engineering" of the Hasso Plattner Institute, and the Federal Ministry of Education and Research (BMBF), Germany (mdViPro, 01IS18092).

References

[AG16] ALPEROVICH A., GOLDLUECKE B.: A Variational Model for Intrinsic Light Field Decomposition. In *Proc. Asian Conference on Computer Vision* (2016), Springer, Cham, pp. 66–82. doi:10.1007/978-3-319-54187-7_5. 2

[AS17] ALAIN M., SMOLIC A.: Light Field Denoising by Sparse 5D Transform Domain Collaborative Filtering. In *Proc. IEEE Workshop on Multimedia Signal Processing* (2017), IEEE, pp. 1–6. doi:10.1109/MMSP.2017.8122232. 2

[ASC*14] AYDIN T. O., STEFANOSKI N., CROCI S., GROSS M., SMOLIC A.: Temporally Coherent Local Tone Mapping of HDR Video. *ACM Trans. Graph. 33*, 6 (2014), 196:1–196:13. doi:10.1145/2661229.2661268. 1, 2

[AZJ*15] AO H., ZHANG Y., JARABO A., MASIA B., LIU Y., GUTIERREZ D., DAI Q.: Light Field Editing Based on Reparameterization. In *Proc. Pacific Rim Conference on Multimedia* (2015), Springer, Cham, pp. 601–610. doi:10.1007/978-3-319-24075-6_58. 3

[BNTS07] BOUSSEAU A., NEYRET F., THOLLOT J., SALESIN D.: Video Watercolorization Using Bidirectional Texture Advection. *ACM Trans. Graph. 26*, 3 (2007), 104:1–104:7. doi:10.1145/1276377.1276507. 2, 10

[BSM*18] BEIGPOUR S., SHEKHAR S., MANSOURYAR M., MYSZKOWSKI K., SEIDEL H.-P.: Light-Field Appearance Editing based on Intrinsic Decomposition. *Journal of Perceptual Imaging 1*, 1 (2018), 10502–1–10502–15. doi:10.2352/J.Percept.Imaging.2018.1.1.010502. 2

[BSPP13] BONNEEL N., SUNKAVALLI K., PARIS S., PFISTER H.: Example-Based Video Color Grading. *ACM Trans. Graph. 32*, 4 (2013), 39:1–39:12. doi:10.1145/2461912.2461939. 1, 2, 6

[BST*14] BONNEEL N., SUNKAVALLI K., TOMPKIN J., SUN D., PARIS S., PFISTER H.: Interactive Intrinsic Video Editing. *ACM Trans. Graph. 33*, 6 (2014), 197:1–197:10. doi:10.1145/2661229.2661253. 1

[BT17] BARRON J. T., TSAI Y.-T.: Fast Fourier Color Constancy. In *Proc. IEEE CVPR* (2017), IEEE, pp. 886–894. doi:10.1109/CVPR.2017.735. 1

[BTS*15] BONNEEL N., TOMPKIN J., SUNKAVALLI K., SUN D., PARIS S., PFISTER H.: Blind Video Temporal Consistency. *ACM Trans. Graph. 34*, 6 (2015), 196:1–196:9. doi:10.1145/2816795.2818107. 1, 3, 6, 7, 10

[BTS*17] BONNEEL N., TOMPKIN J., SUN D., WANG O., SUNKAVALLI K., PARIS S., PFISTER H.: Consistent Video Filtering for Camera Arrays. *Comput. Graph. Forum 36*, 2 (2017). doi:10.1111/cgf.13135. 3

[DAC10] DÍEZ-AJENJO M. A., CAPILLA P.: Spatio-temporal Contrast Sensitivity in the Cardinal Directions of the Colour Space. A Review. *Journal of Optometry 3*, 1 (2010), 2–19. doi:10.3921/joptom.2010.2. 7

[DBH19] DELANOY J., BOUSSEAU A., HERTZMANN A.: Video Motion Stylization by 2D Rigidification. In *ACM/EG Expressive Symposium* (2019), Kaplan C. S., Forbes A., DiVerdi S., (Eds.), The Eurographics Association. doi:10.2312/exp.20191072. 5

[DBZY15] DONG X., BONEV B., ZHU Y., YUILLE A. L.: Region-based Temporally Consistent Video Post-processing. In *Proc. CVPR* (2015), IEEE Computer Society, pp. 714–722. doi:10.1109/CVPR.2015.7298671. 3

[DFKE07] DABOV K. ., FOI A. ., KATKOVNIK V. ., EGIAZARIAN K. .: Image denoising by sparse 3-d transform-domain collaborative filtering. *Trans. Img. Proc. 16*, 8 (Aug. 2007), 2080–2095. URL: https://doi.org/10.1109/TIP.2007.901238, doi:10.1109/TIP.2007.901238. 4

[EMU17] EILERTSEN G., MANTIUK R. K., UNGER J.: A comparative review of tone-mapping algorithms for high dynamic range video. *Comput. Graph. Forum 36*, 2 (May 2017), 565–592. URL: https://doi.org/10.1111/cgf.13148, doi:10.1111/cgf.13148. 2

[FBK15] FORTUN D., BOUTHEMY P., KERVRANN C.: Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding 134* (2015), 1–21. doi:10.1016/j.cviu.2015.02.008. 2

[FG17] FRIGO O., GUILLEMOT C.: Epipolar Plane Diffusion: An Efficient Approach for Light Field Editing. In *Proc. British Machine Vision Conference* (2017), BMVA Press, pp. 1–13. URL: https://hal.archives-ouvertes.fr/hal-01591525. 3

[FL11] FARBMAN Z., LISCHINSKI D.: Tonal Stabilization of Video. *ACM Trans. Graph. 30*, 4 (2011), 89:1–89:10. doi:10.1145/2010324.1964984. 1, 2

[FLJ*14] FIŠER J., LUKÁČ M., JAMRIŠKA O., ČADÍK M., GINGOLD Y., ASENTE P., SÝKORA D.: Color Me Noisy: Example-Based Rendering of Hand-Colored Animations with Temporal Noise Control. *Computer Graphics Forum 33*, 4 (2014), 1–10. doi:10.1111/cgf.12407. 5

[GEB16] GATYS L. A., ECKER A. S., BETHGE M.: Image Style Transfer Using Convolutional Neural Networks. In *Proc. IEEE CVPR* (2016), pp. 2414–2423. doi:10.1109/CVPR.2016.265. 1

[GEZ*17] GARCES E., ECHEVARRIA J. I., ZHANG W., WU H., ZHOU K., GUTIERREZ D.: Intrinsic Light Field Images. *Comput. Graph. Forum 36*, 8 (2017), 589–599. doi:10.1111/cgf.13154. 2

[GGvdW12] GIJSENIJ A., GEVERS T., VAN DE WEIJER J.: Improving Color Constancy by Photometric Edge Weighting. *IEEE Trans. Pat. Anal. Mach. Intel. 34*, 5 (2012), 918–929. doi:10.1109/TPAMI.2011.197. 7

[JAFF16] JOHNSON J., ALAHI A., FEI-FEI L.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proc.*

ECCV (2016), Springer, Cham, pp. 694–711. doi:10.1007/978-3-319-46475-6_43. 1, 5, 7, 10

[JHG*17] JOHANNSEN O., HONAUER K., GOLDLUECKE B., ALPEROVICH A., BATTISTI F., BOK Y., BRIZZI M., CARLI M., CHOE G., DIEBOLD M., GUTSCHE M., JEON H., KWEON I. S., PARK J., PARK J., SCHILLING H., SHENG H., SI L., STRECKE M., SULC A., TAI Y., WANG Q., WANG T., WANNER S., XIONG Z., YU J., ZHANG S., ZHU H.: A Taxonomy and Evaluation of Dense Light Field Depth Estimation Algorithms. In *Proc. IEEE CVPR Workshops* (2017), IEEE, pp. 1795–1812. doi:10.1109/CVPRW.2017.226. 2

[JMG11] JARABO A., MASIA B., GUTIERREZ D.: Efficient Propagation of Light Field Edits. In *Proc. Ibero-American Symposium in Computer Graphics* (2011), pp. 75–80. 3

[JPC*15] JEON H.-G., PARK J., CHOE G., PARK J., BOK Y., TAI Y.-W., KWEON I.: Accurate Depth Map Estimation from a Lenslet Light Field Camera. In *Proc. IEEE CVPR* (2015), IEEE, pp. 1547–1555. doi:10.1109/CVPR.2015.7298762. 2

[KBK*13] KAZHDAN M. M., BURNS R. C., KASTHURI B., LICHTMAN J., VOGELSTEIN R. J., VOGELSTEIN J. T.: *Gradient-Domain Processing for Large EM Image Stacks*. Tech. rep., arXiv.org, 2013. URL: http://arxiv.org/abs/1310.0041. 4

[LHW*18] LAI W.-S., HUANG J.-B., WANG O., SHECHTMAN E., YUMER E., YANG M.-H.: Learning Blind Video Temporal Consistency. In *Proc. ECCV* (2018), Springer, Cham, pp. 170–185. doi:10.1007/978-3-030-01267-0_11. 1, 2, 3, 6, 7, 10

[LHY18] LIU N., HAN J., YANG M.-H.: PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection. In *PRoc. IEEE/CVF CVPR* (2018), IEEE, pp. 3089–3098. doi:10.1109/CVPR.2018.00326. 5

[LWA*12] LANG M., WANG O., AYDIN T., SMOLIC A., GROSS M.: Practical Temporal Consistency for Image-based Graphics Applications. *ACM Trans. Graph. 31*, 4 (2012), 34:1–34:8. doi:10.1145/2185520.2185530. 1, 3

[LXT17] LU C., XIAO Y., TANG C.: Real-Time Video Stylization Using Object Flows. *IEEE Trans. Vis. Comput. Graphics 24*, 6 (2017), 2051–2063. doi:10.1109/TVCG.2017.2700470. 2

[MV12] MITRA K., VEERARAGHAVAN A.: Light Field Denoising, Light Field Superresolution and Stereo Camera Based Refocussing using a GMM Light Field Patch Prior. In *Proc. IEEE CVPR Workshops* (2012), IEEE, pp. 22–28. doi:10.1109/CVPRW.2012.6239346. 2

[MZRT16] MEKA A., ZOLLHÖFER M., RICHARDT C., THEOBALT C.: Live Intrinsic Video. *ACM Trans. Graph. 35*, 4 (2016). doi:10.1145/2897824.2925907. 1, 2

[NB17] NGUYEN R. H. M., BROWN M. S.: Why you should forget luminance conversion and do something better. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017* (2017), pp. 5920–5928. URL: https://doi.org/10.1109/CVPR.2017.627, doi:10.1109/CVPR.2017.627. 5

[OEE*18] OVERBECK R. S., ERICKSON D., EVANGELAKOS D., PHARR M., DEBEVEC P.: A System for Acquiring, Processing, and Rendering Panoramic Light Field Stills for Virtual Reality. *ACM Trans. Graph. 37*, 6 (2018), 197:1–197:15. doi:10.1145/3272127.3275031. 2

[Par08] PARIS S.: Edge-Preserving Smoothing and Mean-Shift Segmentation of Video Streams. In *Proc. ECCV* (2008), pp. 460–473. doi:10.1007/978-3-540-88688-4_34. 1, 2

[PHK11] PARIS S., HASINOFF S. W., KAUTZ J.: Local Laplacian Filters: Edge-aware Image Processing with a Laplacian Pyramid. *ACM Trans. Graph. 30*, 4 (2011), 68:1–68:12. doi:10.1145/2010324.1964963. 6, 7

[PPTM*16] PERAZZI F., PONT-TUSET J., MCWILLIAMS B., GOOL L. V., GROSS M., SORKINE-HORNUNG A.: A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *Proc. IEEE CVPR* (2016), IEEE, pp. 724–732. doi:10.1109/CVPR.2016.85. 7

[Qia99] QIAN N.: On the Momentum Term in Gradient Descent Learning Algorithms. *Neural Networks 12*, 1 (1999), 145–151. doi:10.1016/S0893-6080(98)00116-6. 5

[RDB18] RUDER M., DOSOVITSKIY A., BROX T.: Artistic Style Transfer for Videos and Spherical Images. *International Journal of Computer Vision 126*, 11 (2018), 1199–1219. doi:10.1007/s11263-018-1089-z. 2

[SBZ*18] SHEKHAR S., BEIGPOUR S., ZIEGLER M., CHWESIUK M., PALEN D., MYSZKOWSKI K., KEINERT J., MANTIUK R., DIDYK P.: Light-field intrinsic dataset. In *BMVC* (2018), BMVA Press, p. 120. 10

[SDRJ18] SCHILLING H., DIEBOLD M., ROTHER C., JÄHNE B.: Trust Your Model: Light Field Depth Estimation With Inline Occlusion Handling. In *Proc. IEEE CVPR* (2018), IEEE, pp. 4530–4538. doi:10.1109/CVPR.2018.00476. 2

[TDV19] TASSANO M., DELON J., VEIT T.: An Analysis and Implementation of the FFDNet Image Denoising Method. *Image Processing On Line 9* (2019), 1–25. doi:10.5201/ipol.2019.231. 6

[TM98] TOMASI C., MANDUCHI R.: Bilateral filtering for gray and color images. In *Proceedings of the Sixth International Conference on Computer Vision* (1998), ICCV '98, IEEE Computer Society, pp. 839–. URL: http://dl.acm.org/citation.cfm?id=938978.939190. 4

[VA08] VAISH V., ADAMS A.: The (new) stanford light field archive. http://lightfield.stanford.edu/lfs.html, 2008. 2, 10

[Web96] WEBER E.: *E.H. Weber On The Tactile Senses (2nd Edition)*. Erlbaum (UK) Taylor & Francis, 1996. 5

[Wei74] WEINSTOCK R.: *Calculus of Variations: With Applications to Physics and Engineering*. Dover books on advanced mathematics. Dover Publications, 1974. 5

[WG14] WANNER S., GOLDLUECKE B.: Variational Light Field Analysis for Disparity Estimation and Super-Resolution. *IEEE Trans. Pat. Anal. Mach. Intel. 36*, 3 (2014), 606–619. doi:10.1109/TPAMI.2013.147. 2

[WMJ*17] WU G., MASIA B., JARABO A., ZHANG Y., WANG L., DAI Q., CHAI T., LIU Y.: Light Field Image Processing: An Overview. *IEEE Journal of Selected Topics in Signal Processing 11*, 7 (2017), 926–954. doi:10.1109/JSTSP.2017.2747126. 2, 3

[WSS18] WANG W., SHEN J., SHAO L.: Video Salient Object Detection via Fully Convolutional Networks. *IEEE Transactions on Image Processing 27*, 1 (2018), 38–49. doi:10.1109/TIP.2017.2754941. 5, 6

[WWYL18] WEI C., WANG W., YANG W., LIU J.: Deep Retinex Decomposition for Low-Light Enhancement. In *Proc. British Machine Vision Conference* (2018), BMVA Press, pp. 155:1–155:12. URL: http://bmvc2018.org/contents/papers/0471.pdf. 5, 10

[YCC16] YAO C., CHANG C., CHIEN S.: Example-based video color transfer. In *Proc. IEEE International Conference on Multimedia and Expo* (2016), IEEE, pp. 1–6. doi:10.1109/ICME.2016.7552926. 1

[YCC17] YAO C.-H., CHANG C.-Y., CHIEN S.-Y.: Occlusion-aware Video Temporal Consistency. In *Proc. International Conference on Multimedia* (2017), ACM, pp. 777–785. doi:10.1145/3123266.3123363. 1, 3

[ZIE16] ZHANG R., ISOLA P., EFROS A. A.: Colorful Image Colorization. In *Proc. ECCV* (2016), Springer, Cham, pp. 649–666. doi:10.1007/978-3-319-46487-9_40. 6, 7

[ZZZ17] ZHANG K., ZUO W., ZHANG L.: FFDNet: Toward a Fast and Flexible Solution for CNN based Image Denoising. *IEEE Transactions on Image Processing 27*, 9 (2017), 4608–4622. doi:10.1109/TIP.2018.2839891. 4, 6

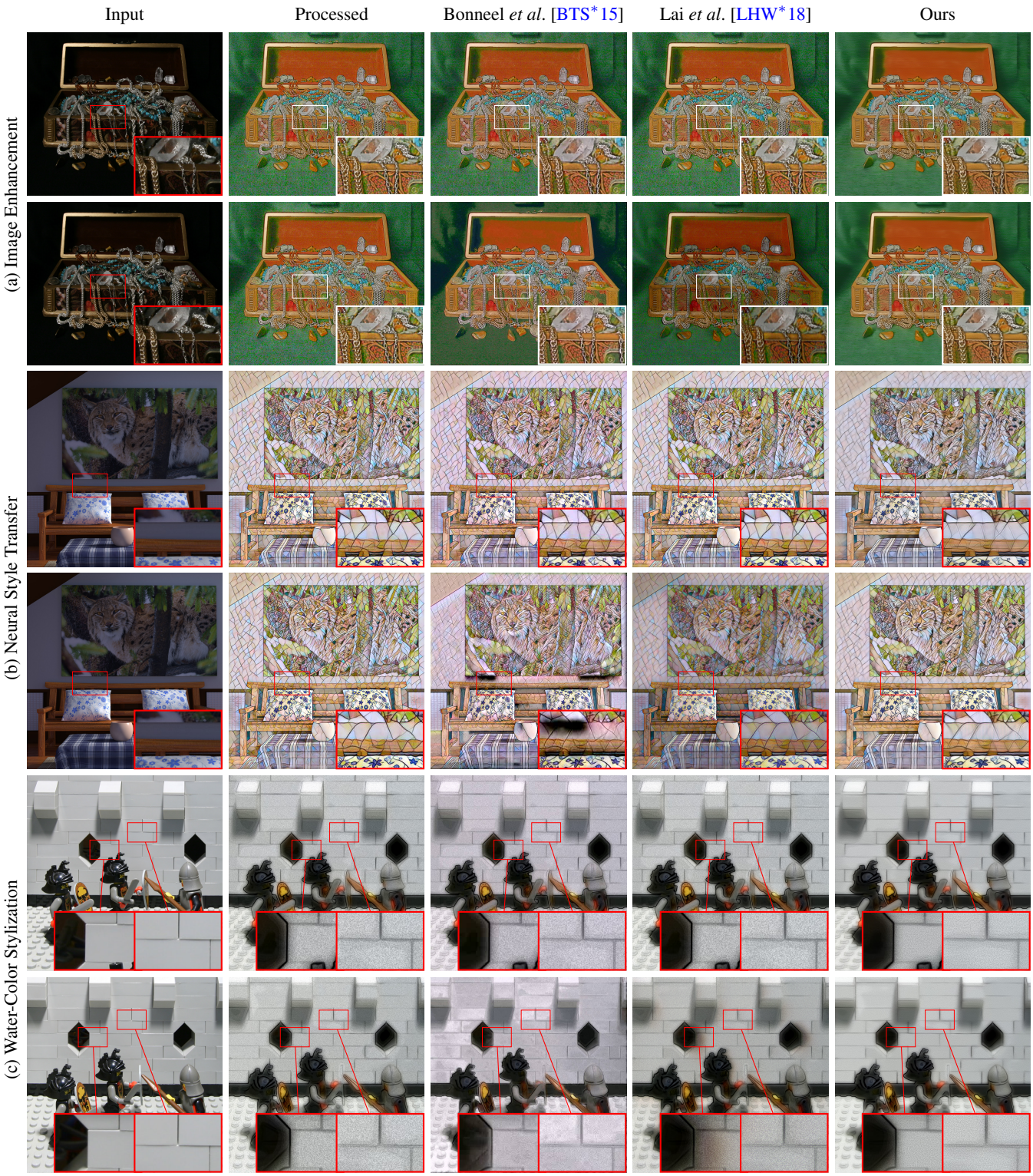


Figure 12: Comparison of our light-field consistency filtering technique with previous methods by applying per-frame (a) Low-light image enhancement by Chen *et al*. [WWYL18] (b) *Neural Style Transfer* by Johnson *et al*. [JAFF16] (c) Water-color stylization with pigment dispersion as proposed by Bousseau *et al*. [BNTS07]. Please refer to supplementary video for better visualization. Input light-fields are taken from the work of Shekhar *et al*. [SBZ*18] and the Stanford light-field archive [VA08].