Burst Reflection Removal using Reflection Motion Aggregation Cues

B H Pawan Prasad[1], Green Rosh K S[1],Lokesh R B[1], Kaushik Mitra[2]
[1]Samsung R&D Institute Bangalore, India [2]IIT Madras, Chennai, India
pawan.prasad,greenrosh.ks,lokesh.rb@samsung.com, kmitra@ee.iitm.ac.in

Abstract

Single image reflection removal has attracted lot of interest in the recent past with data driven approaches demonstrating significant improvements. However deep learning based approaches for multi-image reflection removal remains relatively less explored. The existing multi-image methods require input images to be captured at sufficiently different view points with wide baselines. This makes it cumbersome for the user who is required to capture the scene by moving the camera in multiple directions. A more convenient way is to capture a burst of images in a short time duration without providing any specific instructions to the user. A burst of images captured on a hand-held device provide crucial cues that rely on the subtle handshakes created during the capture process to separate the reflection and the transmission layers. In this paper, we propose a multi-stage deep learning based approach for burst reflection removal. In the first stage, we perform reflection suppression on the individual images. In the second stage, a novel reflection motion aggregation (RMA) cue is extracted that emphasizes the transmission layer more than the reflection layer to aid better layer separation. In our final stage we use this RMA cue as a guide to remove reflections from the input. We provide the first real world burst images dataset along with ground truth for reflection removal that can enable future benchmarking. We evaluate both qualitatively and quantitatively to demonstrate the superiority of the proposed approach. Our method achieves $\sim 2dB$ *improvement in PSNR over single image based methods and* $\sim 1dB$ *over multi-image based methods.*

1. Introduction

Many modern day cameras, especially those present in smart phones have shown significant advances in terms of achieving good image quality in different lighting conditions. Low level computational imaging tasks such as image de-noising [7], high dynamic range (HDR) imaging [10] [12], have shown tremendous improvements in the recent past with the advent of data driven approaches. However, high-level computational imaging tasks such as image inpainting, removing obstructions, reflections, shadows, etc., still pose significant challenges in terms of achieving acceptable image quality. More recently, deep learning based approaches have shown tremendous amount of progress in reflection removal [30] [35] [34] [14] [22] compared to the traditional computationally expensive optimization based methods [25] [16] [17] making them a viable choice to be deployed on consumer products such as smart phones.

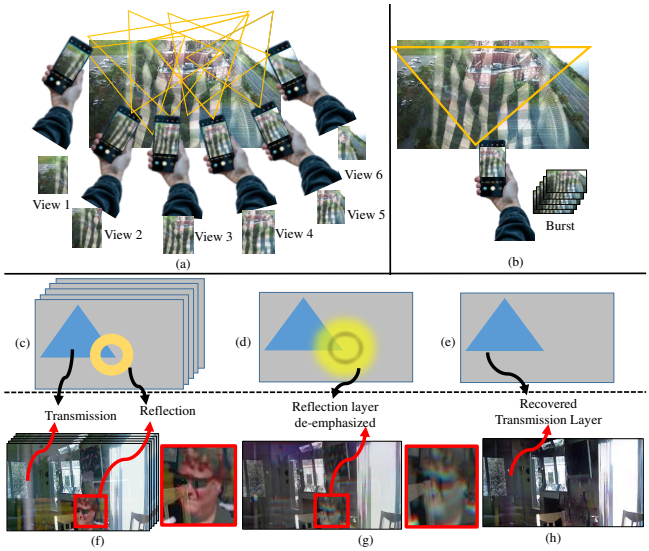


Figure 1. Illustration of the proposed method. **(a)** A typical multi-image capture process (large baseline). **(b)** Our proposed burst capture process for reflection removal that is very convenient to the user. **(c,f)** A typical image containing transmission and reflection layers. **(d,g)** Reflection Motion Aggregation (RMA) cue with reflection layer de-emphasized. **(e,h)** Recovered transmission layer using RMA cue.

An image $\mathbf{I} \in \mathbb{R}^{mxn}$ that is corrupted by reflections can be modeled as a combination of two layers: desirable transmission layer $\mathbf{B} \in \mathbb{R}^{mxn}$ and, the undesirable reflection layer $\mathbf{R} \in \mathbb{R}^{mxn}$ [25]. Removing reflections involves recovering the clear transmission layer $\mathbf{B}$ corrupted by reflections which is significantly ill-posed [17]. Traditional optimization based approaches typically make strong prior assumptions [13] [25] to reduce the ill-posed nature of the

problem. A few other methods impose additional handcrafted constraints such as gradient smoothness [33], ghosting cues [25], sparsity priors [26], relative smoothness [17]. More recently, data driven methods adopt learning based approaches that automatically capture the necessary priors required to recover the clear transmission layer. Reflection removal algorithms, can be further classified depending upon the type of inputs that are used, namely single image based approach or multi-image based approach. While there has been steady progress in single image based deep learning approaches [5], [35], [31], [34], [30], [14], [22], multi-image based deep learning approaches is still relatively less explored. While separating reflection layer and transmission layer using a single input image is severely ill-posed, the multi-image input reduces the ill-posed nature.

Recent methods such as [32] [18] uses multiple images captured at different view points to remove reflections. The method in [32] uses conventional optimization based approach while the more recent [18] deploy a hybrid approach of learning and online optimization to estimate optical flow fields followed by layer reconstruction. This method iteratively optimizes dense motion estimation and layer reconstruction in a coarse to fine manner. Both these methods require a set of images to be captured with different view points to ensure a large baseline between the images. While such methods provide strong cues for reflection removal, it has the following disadvantages. These methods are significantly inconvenient for the end user (see Figure 1(a)) who typically expects an image to be captured instantaneously with a click of a single button. Typically the performance of such a method depends on how well the user follows the capture instructions. Hence, there is a need to devise a method that incorporates the convenience of single image based approaches along with the superior reflection removal capability of multi-image based methods. A burst capture method provides us with this capability.

In this paper we propose a multi-stage burst reflection removal comprising of three stages: (a) Pre-processing, (b) Cue extraction and (c) Guided Reflection removal stages. The first stage individually suppresses the reflections in the burst images. Next we explicitly generate a distinguishing cue from a burst input that is later utilized to guide reflection removal. We propose to extract a cue based on the observation that difference in relative motion between transmission and reflection layers during a burst capture can be leveraged for layer separation. We call this cue as Reflection Motion Aggregation (RMA), the extraction of which is non-trivial and requires careful design consideration. A pictorical representation of the proposed Reflection Motion Aggregation cue is demonstrated in Figure 1. The motion that belong to the reflection layer is identified and aggregated in such a way that the relative smoothness between reflection and transmission layers are made more discernible. The reflection layer is de-emphasized compared to the transmission layer as shown in Figure 1(g).

To summarize, the main contributions are as follows:
(a) A multi-stage deep learning based approach for burst reflection removal that initially performs reflection suppression on individual images followed by extraction of a novel reflection motion aggregation cue that is used to guide the reflection removal. **(b)** The proposed reflection motion aggregation emphasizes the transmission layer more than the reflection layer to further distinguish the two layers that can aid better layer separation. **(c)** We provide a large scale real world burst image dataset with ground truth for reflection removal captured using a smartphone camera in different indoor and outdoor environments for future benchmarking. **(d)** We have performed extensive qualitative and quantitative evaluation along with ablation studies to verify the superiority of the proposed methods over the existing state of the art methods.

2. Related Work

Reflection removal can be categorized into (a) Single Image (b) Multi Image methods [28].
Single Image Reflection Removal: Several past methods use traditional optimization that rely on priors such as gradient sparsity depending upon edges [13], corners [2], layer smoothness [33], different probability distributions to model transmission and reflection layers [17]. A Gaussian mixture model patch based prior with an image formation model comprising of reflection and its spatial shifts was used in [25]. More recently, single image deep learning methods have seen good amount of progress. An end-to-end single image deep learning architecture was proposed in [5]. Usage of perpetual loss was introduced in [35] and a multi scale guided concurrent neural network was proposed in [29]. A non-linear blending model was used to model realistic reflections in [31]. A bi-directional approach where an estimated reflection layer is used to refine the transmission layer was proposed in [34]. An alignment invariant loss function was introduced in [30] that relaxes the constraint of perfect alignment between ground truth and input pairs. A recent method [36] proposes a teacher-student network where the teacher network learns reflection dynamics from multiple view images and a student network is taught to remove reflections from single image during inference.
Multi-Image Reflection Removal: Traditional optimization based approaches use discriminating cues between the two layers present in multiple images. Edge separation cues was used in [16], difference in motion was exploited in [6] [8], dense motion fields using SIFT flow was proposed in [27] and a combination of motion cues and gradient sparsity prior was used in [32]. An approach to remove generic obstructions caused due to glass or fence was proposed in [32]. More recently, Alayrac et.al. [1] proposed to re-

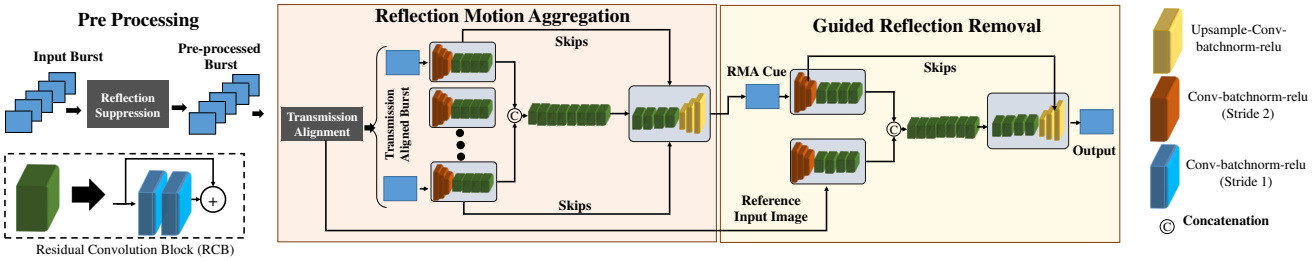


Figure 2. Overview of the proposed Burst Reflection Removal Pipeline. The network architecture of the sub-networks are also shown where second and third stages use similar architecture.

duce the ill-posedness of the problem using multiple input frames. Specifically, they propose a deep neural network named visual centrifuge, which separates different layers of an input video sequence corrupted by obstructions. However this method is proposed as a generic layer separation problem and fails to remove reflections as observed by Liu et.al [18] who proposed a hybrid method to combine optimization with deep learning priors. A more recent method [15] uses the depth ambiguity for edge regeneration to recover transmission layer while [21] uses stereo images to demonstrate layer separation on synthetically generated reflections. Most of these methods rely on images captured at sufficiently different view points by scanning the scene with predefined camera motions making it extremely inconvenient for the end user. Generic layer separation using a burst of images has been recently proposed in [20], however application to reflection removal is relatively less explored.

3. Burst Reflection Removal

In this section, the proposed multi-stage approach is described where we provide details on the network architecture in section 3.1 and loss function used for training the different stages of the network in section 3.2,

3.1. Network Architecture

An overview of the proposed method is shown in Fig. 2. We design the proposed network architecture to leverage relative motion difference between reflection and transmission layers incurred during a burst capture to improve the efficiency of layer separation. We propose to generate a motion aggregate cue to aid reflection removal given a burst of images. The proposed approach consists of 3 stages to generate the final output image. In the first stage, each of the input images are pre-processed to suppress the amount of reflections. These images are passed to the reflection aggregation stage, which aggregates the motion incurred in the reflection layer alone, generating the Reflection Motion Aggregation (RMA) cue image. The RMA image, along with the reference input image (we use middle frame of the burst as reference) are then passed to the guided reflection removal stage, which learns to remove reflection using this cue. Each of these stages are explained in detail in the following subsections.

3.1.1 Single Image Reflection Suppression (SIRS)

The input images are pre-processed to suppress the amount of reflections present. This initial stage aids in the homography based image alignment in the subsequent reflection aggregation stage. We use a modified version of the light weight network architecture proposed in [22]. We use an upsampling operation followed by a convolution in the decoder blocks, instead of using deconvolution, to avoid checkerboard artefacts. We also provide batch normalization and ReLU activation after every convolution operation.

3.1.2 Reflection Motion Aggregation - RMA

The reflection suppressed images are firstly aligned using a homography based image alignment, where we use ORB features [24] and RANSAC [23] to warp the non-reference images to the reference image. The reflection suppression helps in reducing the number of outliers for homography based transmission alignment. The aligned images are then passed to the reflection aggregation stage. The goal of this stage is to aggregate the motion encountered by the reflection layer during the burst duration into a Reflection Motion Aggregate (RMA) image. The reflections that are left out by the reflection suppression stage are accumulated to form the reflection aggregation. Since reflection and transmission layers follow different trajectories during a burst process, it is possible to aggregate the motion in reflection layer alone, without altering the transmission layer. The resultant RMA image will have a higher relative smoothness difference between the transmission and the aggregate reflection layers, compared to the original input. This leads to a better separation of the distributions between the the two layers, which can be leveraged as a cue for better reflection removal in the later stage. In order to handle scenarios where the transmission alignment is not perfect, we deploy a training strategy that can handle misalignments. More details of this training strategy is given in Section 5.

The n outputs from the SIRS stage are warped so that their transmission layers are aligned with that of the refer-

ence image. The transmission aligned images are passed through separate encoders comprising of three convolutional blocks of stride 2 followed by 4 residual dual convolutional blocks (RCBs) as shown in Figure 2. The outputs from the encoders are then concatenated and passed through a sequence of 8 RCBs. The resultant feature maps are then passed through the decoder block consisting of a further 4 RCBs and three upsampling layers to generate the RMA cue image. Each of the upsampling layers perform bilinear upsampling followed by a convolution. We then provide skip connections between all the encoders to the decoders. We also use batch normalization and ReLU activations in every convolutional blocks.

3.1.3 Guided Reflection Removal

In this final stage, the RMA image obtained from the second stage is used as a guide to remove reflections from the reflection suppressed reference image. The RMA image clearly distinguishes between the transmission and reflection layers using motion aggregation. We propose a deep network to use the RMA image as a cue to remove reflections from the reference image. We follow a similar architecture for this stage as that used in RMA stage. We use only 2 encoder blocks to extract features from RMA cue image and the input image respectively.

3.2. Loss Function

We use a combination of contextual loss and pixel losses to train different stages of our pipeline. We introduce small mis-alignments in the inputs to the RMA stage to account for errors in transmission layer alignment that may occur in real-life scenarios. The contextual loss analyses similarity between contextual information and helps in improving convergence [30]. The third stage uses the RMA image as a guide along with input which is also expected to have small mis-alignments, hence we use contextual loss for this stage as well. More details about these loss functions can be found in [30] [19]. We observed that using contextual loss alone results in loss of details and texture. Hence, a pixel level loss is added along with contextual loss as $\mathcal{L}_p = \alpha.\|I_o - I_t\|_1 + \beta.\|I_o - I_t\|_2 + \gamma.\|\Delta I_o - \Delta I_t|_1$. Here Δ represents image gradient. In our experiments, $\alpha = 0.2$, $\beta = 0.2$, $\gamma = 0.4$ were used. The total loss used to train the network is defined as $\mathcal{L} = \mathcal{L}_p + \delta\mathcal{L}_{cx}$ where δ is a weighing parameter and $\mathcal{L}_{cx}$ is the contextual loss. In our experiments, $\delta = 0.8$ was used.

4. Datasets

We use three types of datasets for training and evaluation: (a) Synthetic Burst Reflection dataset, (b) Real world Burst Reflection dataset (ours) and (c) Large Baseline dataset (publicly available).

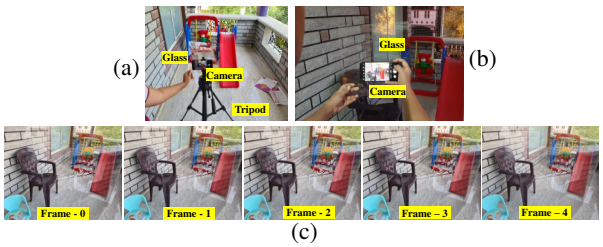


Figure 3. Real World Burst Image Reflection Removal Dataset (BIRR). (a) Data capture setup for the Aligned dataset. (b) Handheld dataset. (c) Sample images of the burst for one set.

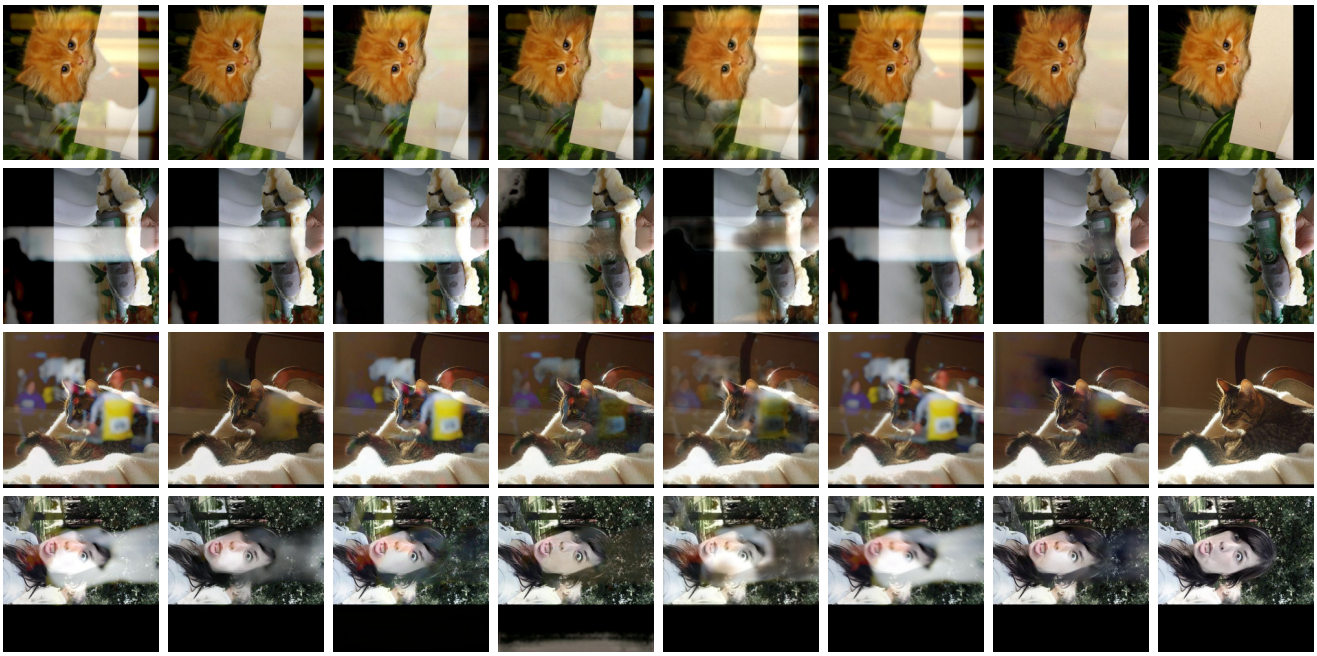
4.1. Synthetic Burst Reflection (SBR) dataset

Learning to remove reflections using burst images requires a large amount of training data. To alleviate this problem, several image formation models have been proposed to generate synthetic data for single image reflection removal [34] [31]. Methods such as [18] extend these models to handle multi-image reflection removal by applying them on video data in a frame-by-frame manner. However, we observed that video data requires extensive pruning to avoid unrealistic scene transitions. Hence, a method is developed to simulate realistic burst capture using a single image. This method generates a synthetic burst with reflection in two stages: a) Synthetic burst generation for transmission and reflection layers; and b) Blending image bursts to generate single burst of images with reflection. Assumptions are made, w.r.t. reflection and transmission layers being planar, static and at different depths from camera plane.

4.1.1 Synthetic Burst Model:

Given images I_t and I_r representing transmission and reflection layers respectively, synthetic image bursts $\mathbf{I_t}$ and $\mathbf{I_r}$ need to be generated. The handshake motion incurred during capture can be represented by the set of motion vectors $M : \{\theta_\mathbf{x}, \theta_\mathbf{y}, \theta_\mathbf{z}, \mathbf{t_x}, \mathbf{t_y}, \mathbf{t_z}\}$, where θ denotes 3 degrees of rotational freedom and $\mathbf{t}$ denotes 3 degrees of translational freedom. Each of the vectors denotes a set sampled from the trajectory followed by the camera during handshake. To generate realistic trajectories, we analyze the work by Kohler et.al. [11] on modeling handshake patterns and observe that a 3^{rd} order polynomial can sufficiently approximate a typical handshake pattern. Next, 6 random polynomials are generated for each of the 6 degrees of freedom from which n points are uniformly sampled to generate M.

For a sufficiently planar scene, [9] showed that M can be approximated using 3 degrees of freedom $M' : \{t_x, t_y, \theta_z\}$ as follows: $\mathbf{t_x} = \mathbf{t_x} - d.sin(\theta_y)$ and $\mathbf{t_y} = \mathbf{t_y} + d.sin(\theta_x)$ [11]. Here z denotes normal axis to camera plane and d denotes distance of image plane from the camera plane. Using the assumption that the reflection and transmission layers are at different depths, we generate two different motion vectors $\{M'_t, M'_r\}$. The synthetic bursts $\{\mathbf{I_t}, \mathbf{I_r}\}$ can then



(a) Input Ref. (b) ERRNet (c) Loc.Aware (d) IBCLN (e) Alayrac (f) Liu.et.al (g) Ours (h) Ground Truth

Figure 4. Qualitative visual comparison against state of the art single image as well as multi-image methods on synthetic dataset.(a) Input burst reference frame, (b) ERRNet [30], (c) Loc.Aware [3], (d) IBCLN [14], (e) Alayrac.et.al [1], (f) Liu.et.al [18] with online optimization, (g) Ours (h) Ground truth. The proposed method is able to suppress reflections to an aesthetically pleasing level (Rows 2,4)

be generated by transforming $\{I_t, I_r\}$ using the homographies represented by $\{M'_t, M'_r\}$.

4.1.2 Synthesizing Reflections:

We use images from PASCAL-VOC dataset [4] for training. In every iteration of training, a pair of bursts $\{\mathbf{I_t}, \mathbf{I_r}\}$ are generated using images $\{I_t, I_r\}$ chosen at random from the dataset. The burst images obtained are then blended using the image formation model proposed by [34]. $\mathbf{I_r}$ is laterally inverted before passing through the image formation model. I_t is chosen as ground truth for training.

4.2. Real World Burst Image Reflection Removal (BIRR) dataset

Our real world burst image reflection removal (BIRR) dataset comprises of a large set of images captured with and without reflections in burst mode. To the best of our knowledge, there is no publicly available large scale burst image dataset for reflection removal. Since training a deep learning method require a large dataset, we capture more than 1000 sets of burst image with at least 5 images per burst. This dataset is split into two categories namely (a) Aligned dataset and (b) Hand-held dataset. We provide the details of each of these sets in the following subsection. We will release the BIRR dataset later upon paper acceptance.

4.2.1 Aligned Dataset:

In order to collect high quality dataset with perfect alignment between input and ground truth transmission, the smartphone is mounted on a tripod to fix the position of the camera. The data collection method for the tripod set is shown in Figure 3. We create real reflections in the scene by using a hand held glass in front of the camera to create different reflection motions for different images in the burst. This captured set of images form the input burst while a single image captured without the hand held glass form the perfectly aligned ground truth transmission for training purpose. We repeat this process to capture a variety of scenarios such as indoor/outdoor, different lighting conditions, different types of glasses, different smartphones, etc to prepare a large set required for training. The input bursts comprises of one reference image and 4 non-reference images, where we choose the middle frame of the burst as reference. We apply synthetic motion model to the non-reference input images using the method described in Section 4.1. The reference image of the burst and the ground truth transmission are perfectly aligned in this dataset. This dataset is mainly used for training purpose.

4.2.2 Hand-Held Dataset:

This dataset comprises of high quality image bursts captured using smartphone cameras in hand-held mode. We capture about 100 different scenes where the glass obstruc-

tion is either hand-held or fixed depending upon the convenience of data capture. We also capture the best possible ground truth image that is required for reference purpose. The reference image of the input burst and the ground truth are not perfectly aligned in this dataset, hence this dataset is only suitable for testing purpose.

4.3. Large Baseline Dataset

We also evaluate the proposed method on the publicly available multi-image dataset for reflection removal [32]. This dataset contains 8 sets of videos using a moving camera to capture the scene at different viewpoints. We sample the video sequence at 5 frames per second and create 5 images per set. This dataset contains large view point differences between the images. In spite of the fact that this is not a burst dataset, we evaluate our method on this dataset as this is only publicly available multi-image dataset for reflection removal.

5. Implementation and Training Details

The proposed method is implemented in Pytorch on a PC with Intel Xeon 6242R with 512GB RAM and Tesla V100 GPU with 32GB memory. The multi-stage architecture comprises of three different networks namely - Single Image Reflection Suppression (SIRS) stage, Reflection Motion Aggregation (RMA) stage and the Guided Reflection Removal stage. Initially, these three networks are independently trained followed by end-to-end transfer learning of the last two stages to achieve convergence.

5.1. Training SIRS stage:

We adopt a strategy similar to [30], [22] where we first train on the single image synthetic datasets generated using the PASCAL VOC dataset [4]. A total of 7000 images with synthetic reflections are generated for training. We train this network for about 1000 epochs using the Adam optimizer with an initial learning rate of 10^{-5} followed by fine tuning on the real world Berkeley dataset [35] for 200 epochs to achieve convergence. We use randomly cropped patches of size 256×256 with scale and rotation augmentations.

5.2. Training RMA and Guided Removal Stages:

Firstly, these two stages are trained independently followed by end-to-end training. The RMA stage is first trained using the synthetic SBR dataset (Section 4.1) with 4000 burst training samples for 1000 epochs using Adam optimizer using a initial learning rate of 10^{-4}. This is followed by fine tuning on the BIRR dataset (Section 4.2) for 500 epochs to achieve convergence. The BIRR aligned dataset is used to generate the ground truth for the RMA stage where we perform weighted averaging of aligned frames. We apply synthetic motion model to the aligned burst frames to generate small mis-alignments to the inputs for training. This strategy of training with mis-aligned input helps the network in handling alignment errors that might have been caused due to incorrect homography estimation. The outputs generated from the RMA stage is then used as input to train the Guided Reflection Removal stage. We follow a similar approach of first training with a synthetic dataset followed by fine tuning on the real dataset. Finally, we perform end-to-end training to fine tune the weights of the network to generate high quality reflection free output. We use a combination of contextual loss and pixel loss as described in Section 3.2 for both these stages.

6. Experimental Results

In this section, we first provide qualitative and quantitative evaluation followed by ablation studies.

6.1. Comparison against State of the Art Methods

We compare the proposed methods on three different datasets (see Section 4 for details). Firstly, we compare both qualitatively and quantitatively on the synthetic datasets followed by the our BIRR dataset. We finally evaluate on the publicly available multi-image dataset that has large view point variations [32]. Please refer to the supplementary material for more sets of comparisons.

6.1.1 Evaluation on the Synthetic SBR Dataset

We first evaluate the proposed method using the synthetic dataset against both single image [30] [14] [3] and multi-image based methods [1] [18]. All the single image methods are trained on synthetic dataset generated using PASCAL-VOC by adapting the method described in [30] [35]. We extend this dataset with synthetic motions to generate the SBR dataset as described in Section 4.1. We use online optimization on this dataset to generate results for [18]. Since, [1] does not release their source code, we implemented their method and trained on this dataset for evaluation purpose. The qualitative comparisons are shown in Figure 4 and the quantitative comparisons in Table 1. As it is evident, our method out-performs the existing single image and multi-image methods both qualitatively as well as quantitatively with at least $1dB$ improvement in PSNR.

6.1.2 Evaluation on the Real World BIRR Dataset

In order to evaluate our method on the real world BIRR dataset, we choose the best performing single image based method on synthetic dataset - ERRNet [30] and also the multi-image based methods [1] [18]. We fine tune both the methods [30] [1] on the BIRR dataset till convergence. We also perform the recommended online optimization for [18] on the BIRR dataset for a fair evaluation. The qualitative evaluation is shown in Figure 5 and the quantitative evaluation in Table 1. While both the multi-image methods [1] [18] are able to suppress the reflection layer to an

(a) Input Ref. (b) ERRNet (c) Alayrac (d) Liu.et.al (e) Ours (f) Ground Truth

Figure 5. Qualitative visual comparison against state of the art single image as well as multi-image methods on our BIRR burst dataset.(a) Input burst reference image, (b) ERRNet [30], (c) Alayrac.et.al [1], (d) Liu.et.al [18], (e) Ours, (f) Ground truth. The proposed method is able to suppress strong reflections better than all other state of the art methods

			SBR Dataset			BIRR Dataset			Large Baseline Dataset		
Method	**Type**	**Time (s)**	**PSNR**	**SSIM**	**NCC**	**PSNR**	**SSIM**	**NCC**	**PSNR**	**SSIM**	**NCC**
ERRNet [30]	Single	0.76	21.89	0.855	0.946	19.51	0.751	0.841	20.49	0.758	0.8418
IBCLN [14]	Single	0.83	21.82	0.836	0.951	18.31	0.737	0.839	18.96	0.735	0.843
Loc.Aware [3]	Single	**0.41**	18.48	0.813	0.910	17.78	0.744	0.941	17.27	0.733	0.855
Liu.et.al [18]	Multi$^{(1)}$	2.35	19.44	0.806	0.874	18.49	0.725	0.814	21.01	0.791	0.852
Liu.et.al [18]	Multi$^{(2)}$	3160	22.78	0.811	0.961	19.59	0.731	0.827	**23.39**	**0.841**	**0.952**
Alayrac.et.al [1]	Multi	1.53	20.91	0.772	0.912	19.13	0.733	0.834	19.71	0.771	0.896
Ours	Multi	1.24	**23.8**	**0.876**	**0.969**	**20.64**	**0.761**	**0.936**	23.03	0.823	0.943

Table 1. Quantitative comparison against on three different datasets. Our method achieves an overall improvement of $1.02dB$ in PSNR on SBR dataset, $1.05dB$ on BIRR dataset, $2dB$ on large baseline dataset over state of the art [18]. Our method is also comparable to a more complex online optimization based method in [18] with $2500\times$ better speed. Here we use two variants of state of the art [18] - Multi$^{(1)}$ Without online optimization, Multi$^{(2)}$ With online optimization. The best results are shown in bold face and the second best is underlined.

extent, they fail to remove the reflection to perceptible levels. This is clearly evident especially in cases of strong reflective components (Rows 3,4). It can be observed that the proposed method is able to suppress the reflection layer better than the state-of-the-art solutions resulting in a cleaner output. The proposed method obtains the best scores on all the metrics of PSNR, SSIM, and NCC.

6.1.3 Evaluation on the Large Baseline Dataset

We also evaluate the proposed method on real video sequences from [32]. This dataset comprises of video sequences captured by moving the camera in order to capture different view points. We prepare a Multi-Image dataset using these video sequences as described in Section 4.3. The proposed method outperforms the single image based methods as well as the multi-image based method [1] and achieves comparable scores against the state of the art method [18]. The qualitative evaluation is shown in Figure 6 and the quantitative evaluation in Table 1. The proposed method is far superior in terms of processing time compared to [18] and is much more suitable for real time applications. The state of the art [18] performs poorly without online optimization. The inference time is measured on a PC with Intel Xeon 6242R with 512GB RAM and an NVIDIA Tesla

(a) Input (b) ERRNet (c) Alayrac (d) Liu.et.al (e) Ours (f) Ground Truth

Figure 6. Qualitative visual comparison against state of the art single image as well as multi-image methods on public Large Baseline dataset (a) Input burst reference (b) ERRNet [30] (c) Alayrac.et.al 2017 [1] (d) Liu.et.al [18] (e) Ours (f) Ground truth. The proposed method is comparable to the state of the art method [18] at significantly lower complexity (see Table 1)

	S1	S2	S3	S4	PSNR	SSIM	NCC
N1	✓[22]				19.23	0.81	0.921
N2	✓[22]	✓		✓	20.04	0.84	0.941
N3		✓	✓	✓	21.81	0.83	0.955
N4	✓[22]	✓	✓	✓	23.8	0.876	0.969
N5	✓[30]	✓	✓	✓	23.96	0.874	0.962
N6	✓[14]	✓	✓	✓	23.12	0.865	0.95
N7	✓[3]	✓	✓	✓	23.62	0.872	0.961

Table 2. Ablation Studies: Network Component Analysis. S1 - SIRS stage, S2 - Transmission Alignment, S3 - Reflection Motion Aggregation, S4 - Guided Reflection Removal

V100 GPU with 32GB memory.

6.2. Ablation studies

We evaluate the proposed pipeline using the SBR dataset the proposed pipeline by considering different design strategies as shown in Table 2. We evaluate the first variation N1 with only the first stage of Single Image Reflection Suppression (SIRS) by providing the reference image of the burst. For the second variation N2, we introduce the transmission alignment as well as guided reflection removal stages. A simple weighted averaging of the transmission aligned images is used instead of RMA stage for this variation. Thirdly for N3, we skip the reflection suppression stage and perform image alignment followed by Reflection Motion Aggregation stage along with the last stage of guided reflection removal to generate the final output. Our next variation N4 is where all the components of the proposed method is introduced. We obtain a gain of about $3dB$ by moving from N2 to N4 where RMA stage is introduced instead of weighted averaging. We also obtain a gain of $1dB$ when the first stage of reflection suppression that helps in improving the homography estimation for improved transmission alignment. Lastly, we evaluate the consideration of the pre-processing step of reflection suppression in N5-N7 by replacing with different state of the art methods [30], [14] and [3]. We see that while [30] provides best results with a higher complexity, [22] provides good results at least complexity.

7. Conclusion

In this paper, a multi-stage deep-learning based approach for removing reflections from a burst of images is proposed. We propose a novel reflection motion aggregation cue for burst reflection removal. We capture a large scale real world dataset that will be publicly released later upon acceptance. Our method achieves more than $2dB$ improvements over single image based methods. For multi-image based methods, our method achieves an overall improvement of $1.02dB$ in PSNR on SBR dataset, $1.05dB$ on BIRR dataset, and $2dB$ on large baseline dataset over state of the art [18]. Our method is also comparable to a more complex online optimization based method in [18] with $2500\times$ better speed.

References

[1] Jean-Baptiste Alayrac, Joao Carreira, and Andrew Zisserman. The visual centrifuge: Model-free layered video representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2457–2466, 2019.

[2] Yun-Chung Chung, Shyang-Lih Chang, Jung-Ming Wang, and Sei-Wang Chen. Interference reflection separation from a single image. In *2009 Workshop on Applications of Computer Vision (WACV)*, pages 1–6. IEEE, 2009.

[3] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Location-aware single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5017–5026, 2021.

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[5] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3238–3247, 2017.

[6] Kun Gai, Zhenwei Shi, and Changshui Zhang. Blind separation of superimposed moving images using image statistics. *IEEE transactions on pattern analysis and machine intelligence*, 34(1):19–32, 2011.

[7] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep burst denoising. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 538–554, 2018.

[8] Xiaojie Guo, Xiaochun Cao, and Yi Ma. Robust separation of reflection from multiple images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2187–2194, 2014.

[9] Ankit Gupta, Neel Joshi, C Lawrence Zitnick, Michael Cohen, and Brian Curless. Single image deblurring using motion density functions. In *European Conference on Computer Vision*, pages 171–184. Springer, 2010.

[10] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017.

[11] Rolf Köhler, Michael Hirsch, Betty Mohler, Bernhard Schölkopf, and Stefan Harmeling. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In *European conference on computer vision*, pages 27–40. Springer, 2012.

[12] Green Rosh KS, Anmol Biswas, Mandakinee Singh Patel, and BH Pawan Prasad. Deep multi-stage learning for hdr with large object motions. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4714–4718. IEEE, 2019.

[13] Anat Levin, Assaf Zomet, and Yair Weiss. Separating reflections from a single image using local features. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[14] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft. Single image reflection removal through cascaded refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3565–3574, 2020.

[15] Tingtian Li, Yuk-Hee Chan, and Daniel PK Lun. Improved multiple-image-based reflection removal algorithm using deep neural networks. *IEEE Transactions on Image Processing*, 30:68–79, 2020.

[16] Yu Li and Michael S Brown. Exploiting reflection change for automatic reflection removal. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2432–2439, 2013.

[17] Yu Li and Michael S Brown. Single image layer separation using relative smoothness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2752–2759, 2014.

[18] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Learning to see through obstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14215–14224, 2020.

[19] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018.

[20] Seonghyeon Nam, Marcus Brubaker, and Michael Brown. Neural image representations for multi-image fusion and layer separation. *arXiv preprint arXiv:2108.01199, 2021*, 2021.

[21] Simon Niklaus, Xuaner Cecilia Zhang, Jonathan T Barron, Neal Wadhwa, Rahul Garg, Feng Liu, and Tianfan Xue. Learned dual-view reflection removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3713–3722, 2021.

[22] BH Pawan Prasad, Lokesh R Boregowda, Kaushik Mitra, Sanjoy Chowdhury, et al. V-desirr: Very fast deep embedded single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2390–2399, 2021.

[23] Rahul Raguram, Jan-Michael Frahm, and Marc Pollefeys. A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus. In *European conference on computer vision*, pages 500–513. Springer, 2008.

[24] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[25] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T Freeman. Reflection removal using ghosting cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3193–3201, 2015.

[26] Tharatch Sirinukulwattana, Gyeongmin Choe, and In So Kweon. Reflection removal using disparity and gradient-sparsity via smoothing algorithm. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1940–1944. IEEE, 2015.

[27] Chao Sun, Shuaicheng Liu, Taotao Yang, Bing Zeng, Zhengning Wang, and Guanghui Liu. Automatic reflection removal using gradient intensity and motion cues. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 466–470, 2016.

[28] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3922–3930, 2017.

[29] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Crrn: Multi-scale guided concurrent reflection removal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2018.

[30] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019.

[31] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3771–3779, 2019.

[32] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T Freeman. A computational approach for obstruction-free photography. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

[33] Qing Yan, Yi Xu, and Xiaokang Yang. Separation of weak reflection from a single superimposed image using gradient profile sharpness. In *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 937–940. IEEE, 2013.

[34] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 654–669, 2018.

[35] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4786–4794, 2018.

[36] Quanlong Zheng, Xiaotian Qiao, Ying Cao, Shi Guo, Lei Zhang, and Rynson WH Lau. Distilling reflection dynamics for single-image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1886–1894, 2021.