Misalignment-Robust Joint Filter for Cross-Modal Image Pairs

Takashi Shibata[1], Masayuki Tanaka[2], Masatoshi Okutomi[2]
[1]NEC corporation, [2]Tokyo Institute of Technology
t-shibata@hw.jp.nec.com, {mtanaka, mxo}@sc.e.titech.ac.jp

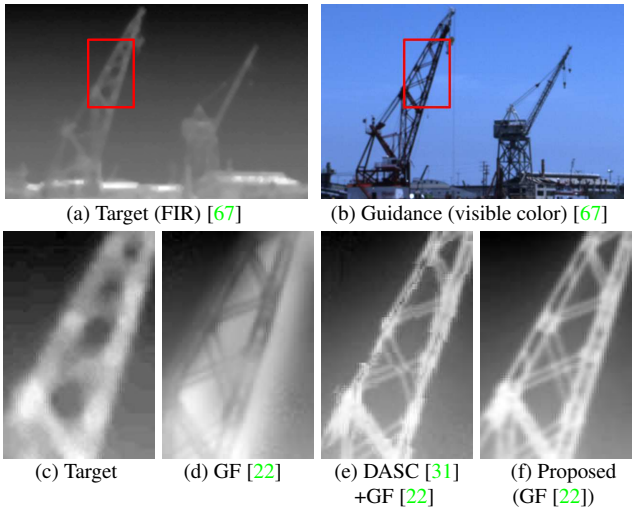
Abstract

Although several powerful joint filters for cross-modal image pairs have been proposed, the existing joint filters generate severe artifacts when there are misalignments between a target and a guidance images. Our goal is to generate an artifact-free output image even from the misaligned target and guidance images. We propose a novel misalignment-robust joint filter based on weight-volume-based image composition and joint-filter cost volume. Our proposed method first generates a set of translated guidances. Next, the joint-filter cost volume and a set of filtered images are computed from the target image and the set of the translated guidances. Then, a weight volume is obtained from the joint-filter cost volume while considering a spatial smoothness and a label-sparseness. The final output image is composed by fusing the set of the filtered images with the weight volume for the filtered images. The key is to generate the final output image directly from the set of the filtered images by weighted averaging using the weight volume that is obtained from the joint-filter cost volume. The proposed framework is widely applicable and can involve any kind of joint filter. Experimental results show that the proposed method is effective for various applications including image denosing, image up-sampling, haze removal and depth map interpolation.

1. Introduction

Recent developments of image sensors and hardware technologies enable the simultaneous capturing of cross-modal image pairs such as visible color and near-infrared (NIR) [50, 69, 43], flash and no-flash [48], visible color and far-infrared (FIR) [45], multi-spectral [7, 61], and visible color and depth [19]. Inspired by these progresses, joint use of the cross-modal image pairs becomes more common in computer vision and pattern recognition applications such as image denosing [62, 70], haze removal [52, 57, 6, 56], image enhancement [33, 69, 59, 17, 18, 8, 58], image up-sampling [42, 47, 40, 15, 21, 37], scene classification [4], pedestrian detection [65, 28], and face recognition [38].

For these applications, joint image filters such as joint bilateral filter (JBF) [48] and guided filter (GF) [22] are effective and commonly used. In general, these filters improve a target image quality by utilizing structural information existing in a guidance image. The vital assumption of these joint filters is that the input image pair is strictly aligned. However, actual cross-modal image pairs usually contain misalignments because these images are captured by different sensors with different view points. The misalignment often generates severe artifacts such as a ghost, a halo and a discontinuity in the filter results. A straight forward approach for reducing these artifacts is to use an aligned guidance obtained by local flow estimation. Many calibration techniques [25, 28, 66] and registration methods [54, 31] for the cross-modal image pairs have been proposed. Although these methods can estimate the rough flow between the cross-modal image pairs, accuracy is still insufficient for these joint filters.



(a) Target (FIR) [67] (b) Guidance (visible color) [67]
(c) Target (d) GF [22] (e) DASC [31] +GF [22] (f) Proposed (GF [22])

Figure 1. Up-sampling example of existing and the proposed methods. The misalignment between the cross-modal image pair generates artifacts for the existing methods as shown in (d) and (e). Our method can generate an artifact-free result as shown in (f).

An example of the results on image up-sampling by the existing joint filters is shown in Fig. 1. Here, the target far-infrared (FIR) image and the guidance visible color image are taken from slightly different view points, which causes the misalignment. As shown in Fig. 1 (d), GF [22] generates the ghost artifacts due to the misalignment. Although the ghost artifacts can be reduced using the aligned guid-

ance obtained by applying dense adaptive self-correlation descriptor (DASC) [31], which is one of the state-of-the-art registration methods for the cross-modal image pairs, the registration error still remains and severely degrades the result as shown in Fig. 1 (e).

We propose a novel misalignment-robust joint filter for cross-modal image pair based on 1) **weight-volume-based image composition** and 2) **joint-filter cost volume**. The proposed method can generate an artifact-free filtered output image by weighted averaging using the weight volume without estimating local flow. This weight volume is obtained from the joint-filter cost volume. The result by the proposed method is shown in Fig. 1 (f). The artifact-free output image can be generated by the proposed method. The proposed framework is widely applicable and can involve any kind of joint filter such as GF [22], mutual-structure for joint filtering (MSJF) [55], and JBF [48]. The applications of the proposed method includes image denosing, image up-sampling, haze removal, and depth map interpolation.

2. Related works

2.1. Joint filter for cross-modal image pair

Various joint filters for cross-modal image pairs and their applications have been proposed. Petschnigg *et al*. proposed the joint bilateral filter (JBF) [48] for flash/no-flash photography. The JBF is also useful for image up-sampling [34] and image fusion [16]. The guided filter (GF) [22] and its extended version [68, 11] are applicable for image denoising, image up-sampling and haze removal [23]. The weighted least square filter (WLS) [14] for image enhancement using an NIR image was also presented [70]. Shen *et al*. proposed the mutual-structure for joint filtering (MSJF) [55] which addresses mutual-structure consistency between cross-modal image pair. Recently, the image restoration method via scale-map was also presented by Yan *et al*. [63]. Other joint filters which are applicable for cross-modal image pairs are the non-local means [5] and the joint static and dynamic guidance filter [20].

These existing joint filters generate severe artifacts when there is the misalignment between the cross-modal image pair. The proposed framework potentially makes any existing joint filter robust against the misalignment.

2.2. Local-flow estimation for warped image generation

One of the most common local flow estimation methods for visible image pairs is Horn-Schunk method [26]. The Horn-Schunk method estimates local flow by optimizing an energy which consists of a data term based on sum of square difference (SSD) and a spatial smoothness term. This energy can be optimized by a gradient method because the data term is derivative.

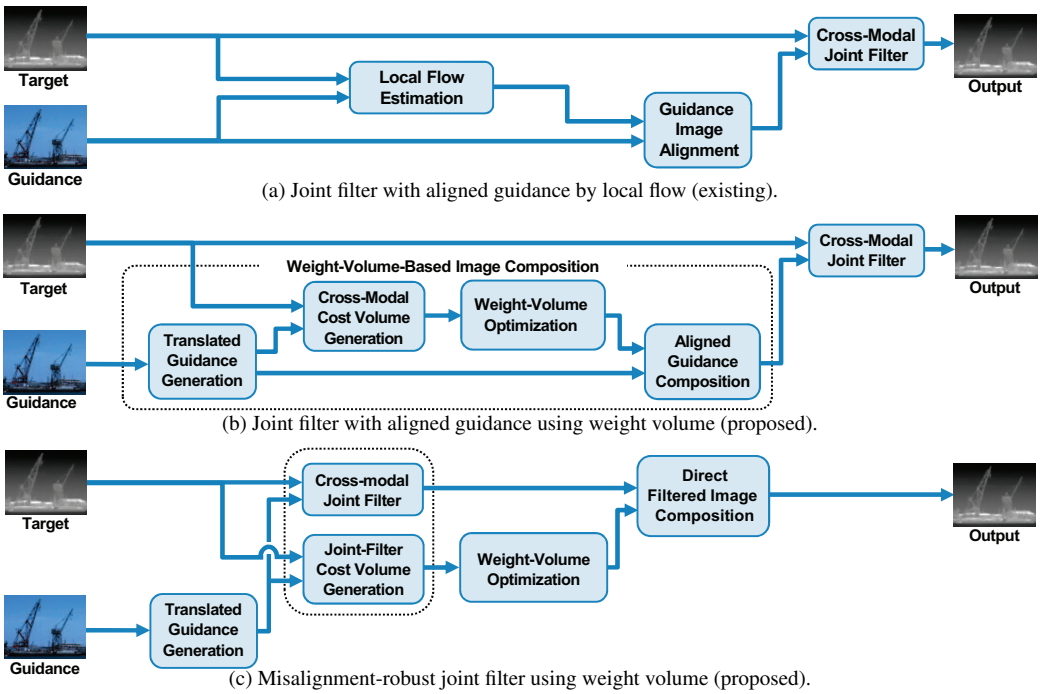
Although SSD is widely used similarity measure for visible image pairs, it is not suitable for cross-modal image pairs because the appearance difference between the cross-modal image pair is more significant than that of the visible ones. To measure the accurate similarity between the cross-modal image pair, various similarity measures have been proposed such as mutual information (MI) [49, 12], normalized cross-correlation (NCC) on the Laplacian pyramid [29], the adaptive normalized cross-correlation (ANCC) [24], and the robust selective normalized cross-correlation (RSNCC) [54]. Recently, Kim *et al*. proposed the dense adaptive self-correlation descriptor (DASC) [31] and its improved version [32] inspired by the local self-similarity descriptor (LSS) [53].

In general, these similarity measures for the cross-modal image pair is non-derivative. Therefore, the gradient method cannot be used to minimize the energy based on these similarity measures. To address this problem, DASC [31] estimates discrete local flow by optimizing the discrete energy originally designed for SIFT flow [41]. On the other hand, many stereo vision algorithms evaluate the cost volume [27, 51, 64, 3] constructed based on the non-derivative similarity such as NCC. The *winner-take-all* is usually used to determine the disparity.

3. Proposed method

In actual situation, cameras for the cross-modal image pair are placed reasonably close to each other, and we can easily obtain the roughly aligned images by performing camera calibration and/or by applying global image registration. However, misalignment still exists and this remaining displacement cannot be modeled by a rigid motion. Our goal is to generate an artifact-free filtered image from misaligned cross-modal image pairs. We propose the misalignment-robust joint filter based on 1) **weight-volume-based image composition** and 2) **joint-filter cost volume**.

The naive approach for this goal is to use aligned guidance obtained by local flow estimation (Fig. 2 (a)). In general, however, the accuracy of the estimated local flow is still insufficient for cross-modal joint filters. In the proposed approach by the weight-volume-based image composition, the aligned guidance is composed by fusing a set of translated guidances using the weight volume (Fig. 2 (b)). The weight-volume-based image composition becomes more powerful scheme by calculating the weight volume from the joint-filter cost volume. As shown in Fig. 2 (c), the proposed misalignment-robust joint filter using the weight volume can compose the final output image by fusing the set of the filtered images and the weight volume without the aligned guidance and the local flow. Our method can remarkably improve the robustness of existing joint filters against misalignment where the displacements are supposed to be within a predefined range.



(a) Joint filter with aligned guidance by local flow (existing).

(b) Joint filter with aligned guidance using weight volume (proposed).

(c) Misalignment-robust joint filter using weight volume (proposed).

Figure 2. Image processing pipelines.

3.1. Joint filter with aligned guidance using weight volume

We first propose the joint filter with aligned guidance by the weight-volume-based image composition. The key is to compose the aligned guidance by fusing the set of the translated guidances using the weight volume. In the proposed approach, the unique local flow is not required to generate the aligned guidance.

As shown in Fig. 2 (b), this approach first generates the set of the translated guidances. Then, a cross-modal cost volume is calculated by measuring the similarity between the target image and the set of the translated guidances. Next, we obtain the weight volume which represents the confidence of each translated guidance. The aligned guidance is composed by weighted averaging of the set of the translated guidances using the weight volume. Finally, the filtered image is generated by applying an existing cross-modal joint filter to the target image with the aligned guidance.

Translated guidance generation: First of all, the set of the translated guidances are generated by translating the original guidance image. Let $\mathbf{t} = (t_1, \cdots, t_i, \cdots, t_N)^T$ and $\mathbf{g} = (g_1, \cdots, g_i, \cdots, g_N)^T$ be the vectorized target image and the vectorized original guidance image, where i is the pixel index. The set of the translated guidances are given by $\{\mathbf{H}_1\mathbf{g}, \cdots, \mathbf{H}_k\mathbf{g}, \cdots, \mathbf{H}_K\mathbf{g}\}$, where $\mathbf{H}_k$ is the matrix operator for k-th labeled translation vector which is corresponding to k-th fixed flow. Here, the range of the translation is corresponding to the horizontal and vertical ranges of the local flow. This range (e.g. 7 [pix] $\times$ 7 [pix]) is the one of the setting parameters. The local translation range depends on the resolution of the input image pair.

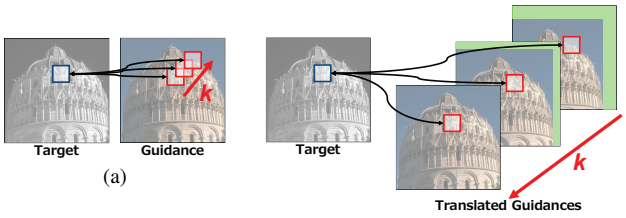


Figure 3. Local flow evaluation processes. (a) Evaluate all possible flows (k) for a fixed pixel. This is repeated with different pixels. (b) Evaluate a fixed flow for all pixels. This is repeated with different flows (k). (proposed method).

Note that the proposed method can deal with the non-uniform local flow like many existing local flow estimation, *e.g.* Horn-Schunk [26], DASC [31] and SIFT flow [41]. As shown in Fig. 3 (a), these existing methods evaluate all possible flows (k) for a fixed pixel. This is repeated with different pixels. On the other hand, as shown in Fig. 3 (b), the proposed method evaluates a fixed flow for all pixels. This is repeated with different flows (k). In other words, to evaluate the local flow, the proposed method translates the guidance images, whereas the existing algorithms translate the one side patch with opposite direction. In this sense, the proposed method can evaluate the local flow like the existing local flow estimation methods [26, 31, 41].

Cross-modal cost volume generation: Next, we evaluate the cross-modal cost volume. Cost volume approach is widely used for estimating disparity in stereo vision [27, 51] because it can deal with various similarity measures with simple and effective implementation. The proposed weight-

volume-based image composition is inspired by the successes of the cost-volume in stereo vision.

Let $\mathbf{C} = [\mathbf{c}_1, \cdots, \mathbf{c}_k, \cdots, \mathbf{c}_K]$ be the cross-modal cost volume, where $\mathbf{c}_k$ represents k-th vectorized cross-modal cost map. Each element of the cross-modal cost volume $\mathbf{C}$ is given by $c_{i,k}$, where i is the pixel index. The cross-modal cost map is calculated from the similarity measure between the target image $\mathbf{t}$ and the translated guidance $\mathbf{H}_k\mathbf{g}$ as

$$\mathbf{c}_k = dist(\mathbf{t}, \mathbf{H}_k\mathbf{g}), \tag{1}$$

where k is the label corresponding to the translation vector, and $dist(\cdot,\cdot)$ is the similarity measure such as DASC [31] and NCC [29].

Weight volume optimization: Then, the weight volume is calculated from the cross-modal cost volume $\mathbf{C}$. The variable range of the cross-modal cost volume $\mathbf{C}$ significantly depends on the similarity measures. To correct this range, we introduce the weight volume $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_k, \cdots, \mathbf{w}_K]$ by normalizing the cross-modal cost volume $\mathbf{C}$ while considering the spatial-smoothness constraint and the label-sparseness constraint. Here, each element of the cross-modal cost volume $\mathbf{W}$ is given by $w_{i,k}$. The details of the weight volume is described in the next section.

Aligned guidance composition: The aligned guidance $\mathbf{y}$ is composed by averaging the set of the translated guidances $\{\mathbf{H}_1\mathbf{g}, \cdots, \mathbf{H}_k\mathbf{g}, \cdots, \mathbf{H}_K\mathbf{g}\}$ as

$$\mathbf{y} = \sum_k \mathrm{diag}(\mathbf{w}_k)\mathbf{H}_k\mathbf{g}. \tag{2}$$

Finally, the output image $\mathbf{z}$ is generated by applying an existing cross-modal joint filter [22, 55, 48] to the target image $\mathbf{t}$ with the aligned guidance $\mathbf{y}$. Note that the proposed weight volume $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_k, \cdots, \mathbf{w}_K]$ can be considered as a probability for each translation vector labeled by k. In this sense, the proposed weight-volume-based image composition in Eq. (2) can be also considered as the expected value of the set of the translated guidances based on the *probability* $\mathbf{W}$. As described in Sec. 3.4, the probabilistic approach can reduce the artifacts in the aligned guidance, which directly generates the poor result in the final output.

3.2. Weight volume optimization

The weight volume $\mathbf{W}$ is obtained by optimizing the energy which consists of three terms: 1) the fidelity term, 2) the spatial-smoothness constraint, and 3) the label-sparseness constraint. The proposed energy for the weight-volume optimization is given by

$$\begin{aligned} F(\mathbf{W}) &= F_d(\mathbf{W}) + \eta F_{ss}(\mathbf{W}) + \gamma F_{ls}(\mathbf{W}), \\ s.t.\ \ &^\forall i,\ \sum_k w_{i,k} = 1,\ \ w_{i,k} \geq 0, \end{aligned} \tag{3}$$

where k is the label corresponding to each translation vector, i is the pixel index, $w_{i,k}$ is the element of the weight-volume element for the k-th label and i-th pixel, $F_d(\mathbf{W})$ is the data term, $F_{ss}(\mathbf{W})$ is the spatial-smoothness constraint, and $F_{ls}(\mathbf{W})$ is the label-sparseness constraint. Here, η (typically 10 to 100) and γ (typically 0 to 1) are the parameters to control the strength of the spatial-smoothness constraint and the label-sparseness constraint.

The data term penalizes the residual between the weight volume $\mathbf{W}$ and the normalized cross-modal cost volume $\tilde{\mathbf{W}}$ as

$$F_d(\mathbf{W}) = \sum_k \sum_i ||\mathbf{W} - \tilde{\mathbf{W}}||_F^2, \tag{4}$$

where $\tilde{w}_{i,k}$ is the element of the normalized cross-modal cost volume $\tilde{\mathbf{W}}$, and is obtained by applying soft-max function to the element of the cross-modal cost volume $c_{i,k}$ as

$$\tilde{w}_{i,k} = \frac{\exp\left[-\beta c_{i,k}\right]}{\sum_k \exp\left[-\beta c_{i,k}\right]}, \tag{5}$$

where β (typically 1 to 10) is the parameter depending on the range of the cross-modal cost volume element $c_{i,k}$.

To generate the spatially smooth output image, we introduce the spatial-smoothness constraint for the weight volume as

$$F_{ss}(\mathbf{W}) = ||\mathbf{D}_h\mathbf{W}||_F^2 + ||\mathbf{D}_v\mathbf{W}||_F^2, \tag{6}$$

where $\mathbf{D}_h$ and $\mathbf{D}_v$ is horizontal and vertical derivative operator matrix.

The output image is composed by averaging the set of the translated guidances using the weight volume. In general, artifacts such as discontinuity, blur, and ghost tend to be generated by the noisy small weight volume element. To reduce these artifacts, we introduce the label-sparseness constraint for the weight-volume as

$$F_{ls}(\mathbf{W}) = \sum_i \|w_{i,k}\|_{p,k} = \sum_i \left(\sum_k |w_{i,k}|^p \right)^{1/p}, \tag{7}$$

where $\|\cdot\|_{p,k}$ denotes L_p norm w.r.t. label k, and $p(>0)$ is the label-sparseness parameter. We set p as 0.5[1].

In the proposed method, the energy $F(\mathbf{W})$ is optimized based on a proximal approach [9, 46]. The weight-volume optimization is the most computationally expensive process, which is proportional to the number of the iteration and the number of the translated guidances. It takes several minutes with non-optimized implementation in MATLAB.

3.3. Misalignment-robust joint filter using weight volume

The weight-volume-based image composition described in Sec. 3.1 can generate the aligned guidance without estimating local flow. However, our goal is to generate an

[1]The label-sparseness constraint can be effectively solved by the proximal mapping approach. Particularly, we can be easily obtained the proximal mapping for $p = 0.5$ by analytical approach [36].

artifact-free filtered image from misaligned cross-modal image pairs. In this sense, it is not necessarily required to generate the aligned guidance for the joint filter. Furthermore, the use of the cross-modal cost volume dose not guarantee to generate the suitable filtered output, because the similarity between the cross-modal image pair does not correspond to the confidence of the filtered output. We introduce the joint-filtered cost volume $\mathbf{C}^{\mathrm{jf}}$ which is directly corresponding to the confidence of the joint filtered image. By using the weight-volume-based image composition and the joint-filter cost volume, the misalignment-robust joint filter can directly generate the final output from the set of the translated guidances without generating the aligned guidance and the unique local flow.

The processing pipeline is shown in Fig. 2 (c). First, the set of the translated guidance is generated. Then, the set of the filtered images and the joint-filter cost volume are generated from the target and the set of the translated guidances. Next, the weight volume is obtained from the joint-filter cost volume. Finally, the final output is directly composed by averaging the set of the filtered images using the weight volume.

Cross-modal joint filter and joint-filter cost volume generation: In the misalignment-robust joint filter, we generate the set of filtered images and the joint-filtered cost volume $\mathbf{C}^{\mathrm{jf}}$ instead of the cross-modal cost volume $\mathbf{C}$. Let $\mathbf{j}_k = (j_1^k, \cdots, j_i^k, \cdots j_N^k)^T$ be the filtered image obtained from the target image $\mathbf{t}$ and the translated guidance $\mathbf{H}_k\mathbf{g}$ by an existing cross-modal joint filters, where k and i are label's and pixel's indexes.

The cross-modal joint filters such as GF [22] generate the filtered image by minimizing the cost pixel-by-pixel, which is designed so that features, *e.g.* structures and textures, of both images are harmonized. For example, the cost function at the i-th pixel for the original GF [22] is

$$E(a_i, b_i) = \sum_{l \in \mathcal{N}_i} \Big(\big((a_i g_l + b_i - t_l) \big)^2 + \varepsilon a_i^2 \Big), \tag{8}$$

where t_i and g_i are the i-th pixel's value of the target image $\mathbf{t}$ and the guidance image $\mathbf{g}$, a_i and b_i are the coefficients for the linear transformation, $\mathcal{N}_i$ is the set of neighboring pixels at the i'th pixel, and ε is the regularization parameter. The first term describes the residual between the target image and the linear-transformed guidance image, whereas the second term represents the regularization.

The joint filter cost volume $\mathbf{C}^{\mathrm{jf}}$ for GF [22] is designed based on the residual term in Eq. (8) as

$$c^{\mathrm{jf}}_{i,k} = \sum_{l \in \mathcal{N}_i} \Big(a_{k,i} (\mathbf{H}_k\mathbf{g})_l + b_{k,i} - t_l) \Big)^2, \tag{9}$$

where $c^{\mathrm{jf}}_{i,k}$ is the element of the joint-filter cost volume at k-th label and i-th pixel, $(\mathbf{H}_k\mathbf{g})_l$ is the k-th labeled translated guidance $\mathbf{H}_k\mathbf{g}$ at the pixel l, $a_{k,i}$ and $b_{k,i}$ are the linear transformation coefficients for $(\mathbf{H}_k\mathbf{g})_l$. Note that the proposed method is the general framework for the cross-modal joint filters. We can apply the proposed joint filter cost volume approach to existing joint filters including MSJF [55], the scale map image restoration [63], the dark flash photography [35], if the cost function of the joint filter is defined pixel-by-pixel. Other examples of the joint-filter cost volume are described in our supplemental material.

The recent studies on image filtering [44, 13] also showed that the classical joint filter such as JBF [48] can be formulated as the cost function minimization by the kernel function. Based on these studies, we can also define the joint-filter cost volume of these classical joint filters.

Direct filtered image composition: After calculating the joint-filtered cost volume $\mathbf{C}^{\mathrm{jf}}$, we calculate the weight volume $\mathbf{W}$ using Eqs. (3) to (7) in the same manner as described in Sec. 3.2. Note that, instead of Eq. (7), the element of the normalized joint-filter cost volume $\tilde{w}_{i,k}$ is given by

$$\tilde{w}_{i,k} = \frac{\exp\big[-\beta c^{\mathrm{jf}}_{i,k}\big]}{\sum_k \exp\big[-\beta c^{\mathrm{jf}}_{i,k}\big]}. \tag{10}$$

The final output $\mathbf{z}$ is directly generated by fusing the set of the joint-filtered images $\{\mathbf{j}_1, \cdots, \mathbf{j}_k, \cdots, \mathbf{j}_K\}$ and the weight volume $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_k, \cdots, \mathbf{w}_K]$ as

$$\mathbf{z} = \sum_k \mathrm{diag}(\mathbf{w}_k)\mathbf{j}_k. \tag{11}$$

As with Eq. (2), the proposed composition in Eq. (11) can be also considered as expectation of the set of the filtered images based on the weight volume as *probability*.

3.4. Comparisons

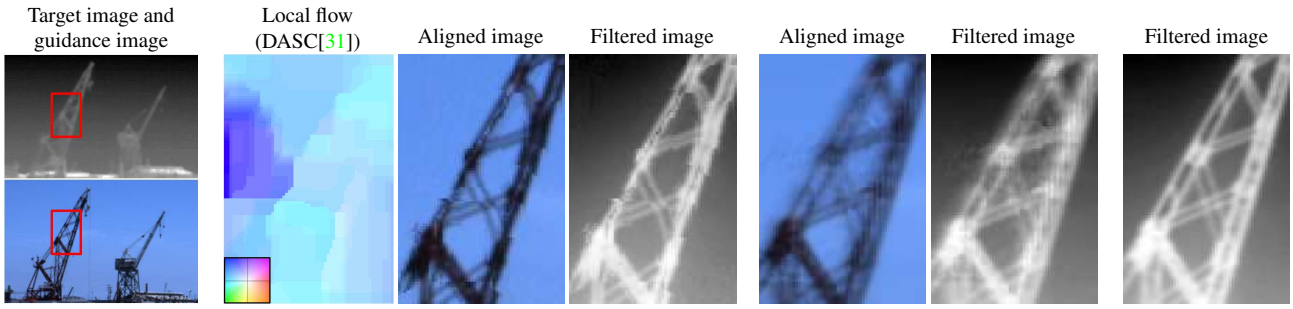
To discuss the effectiveness of the weigh-volume-based image composition and the joint-filter cost volume, we compared with the three approaches shown in Fig. 2.

a) Joint filter with aligned guidance by local flow (Fig. 2 (a)): The local flow was estimated by DASC [31]. GF [22] was used as the cross-modal joint filter.

b) Joint filter with aligned guidance using weight volume (Fig. 2 (b)): DASC [31] and GF [22] were used for the cross-modal cost volume and the cross-modal joint filter.

c) Misalignment-robust joint filter using weight volume (Fig. 2 (c)): GF [22] and the corresponding joint-filtered cost volume in Eq. (9) were used.

An example of the results is shown in Fig. 4. As shown in Fig. 4 (a), although we can estimate the rough flow by DASC [31], the aligned guidance contains the collapsed artifacts due to the subtle alignment error. This artifacts directly generates the poor results in the filtered image. In joint filter with aligned guidance using weight volume, the collapsed artifacts can be reduced as shown in Fig. 4 (b). Although the result shows the effectiveness of the weight-volume-based image composition, the aligned image and the filtered image are still blurred. On the other hand,



(a) Joint filter with aligned guidance by local flow (Fig. 2 (a))
(b) Joint filter with aligned guidance using weight volume (Fig. 2 (b))
(c)Misalignment-robust joint filter (Fig. 2 (c))

Figure 4. Comparison with existing and proposed approaches shown in Fig. 2.

Fig. 4 (c) show that the misalignment-robust joint filter using weight volume can generate the clear image without blurred and collapsed artifacts. These results show that the misalignment-robust joint filter using the weight volume and the joint filter cost volume can generate clearer filtered image than the joint filter with aligned guidance.

4. Experiments

The proposed method has various applications such as image denosing, image up-sampling and depth interpolation using misaligned cross-modal image pairs. In this section, we demonstrate the examples of the performance of the proposed method for each application[2]. Note that almost figures are close-up images due to the space limitations. Whole images and additional results are presented in our supplemental materials.

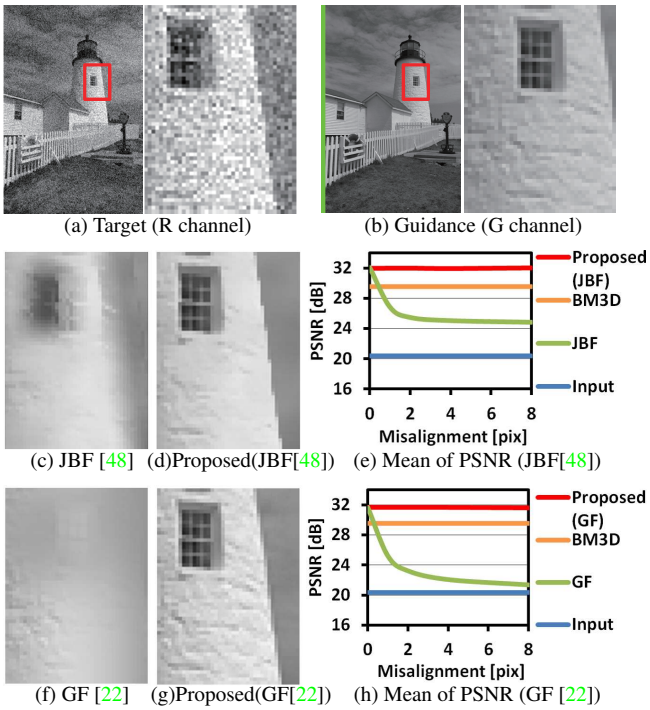
4.1. Image denosing

Robustness against misalignment: We first demonstrate the robustness of the proposed method against the misalignment by image denoising experiments using the simulated images, where the misalignments are artificially synthesized. To examine the robustness against the misalignment accurately, we used the 12 image pairs of the R channel and G channel image from the Kodak color image dataset [39], which are well-aligned. Here, the R channel images were used as the noisy target images obtained by adding Gussian white noise (σ=25), whereas the shifted G channel images were used as the guidance image. The misalignment were synthesized by shifting the original guidance image in the horizontal direction.

An example of the results by JBF [48], GF [22] and the proposed method with these cross-modal joint filters are shown in Fig. 5. As shown in Fig. 5 (c) and (f), the blurred results are generated by naively applying and JBF [48] and GF [22]. On the other hand, Fig. 5 (d) and (g) show that the proposed method can remove the noise without the blur artifacts.

To evaluate the performance of the proposed method quantitatively, we measured peak signal-to-noise ratio (PSNR) between the output results and the ground truth.

[2]The code will be available at http://www.ok.sc.e.titech.ac.jp/res/MMIP.



(a) Target (R channel) (b) Guidance (G channel)
(c) JBF [48] (d)Proposed(JBF[48]) (e) Mean of PSNR (JBF[48])
(f) GF [22] (g)Proposed(GF[22]) (h) Mean of PSNR (GF [22])

Figure 5. Robustness against misalignment between target and guidance images. The guidance (b) and the results (c), (d), (f), and (g) is generated by adding the 8-pixel horizontal shift (green rectangle). (σ=25)

Here, we also evaluated the performance of BM3D [10] as reference. As shown in Fig. 5 (e) and (h), PSNR are dramatically decreased by naively applying JBF [48] and GF [22]. The proposed method can maintain the PSNR while the misalignment exists between the target and the guidance images. The results with another image quality measure such as the structural similarity (SSIM) [60] and the results for other cross-modal joint filters, *e.g.* MSJF [55], are shown in the supplemental material.

Comparison with the existing methods: Next, we evaluated the performance of the proposed method using the natural cross-modal image pairs collected by Brown *et al.* [4]. In our experiment, we selected 40 image pairs of the visible and the NIR images which naturally contains the misalignment. Here, the noisy visible color images obtained by

(a)Target(VIS) (b)Guidance(NIR) (c) Ground truth (d) BM3D [10]

(e) GF [22] (f) DASC [31]+GF [22] (g) Proposed(GF [22])

Figure 6. Visible image denosing guided by NIR image. ($\sigma = 50$)

Table 1. Mean of PSNR and SSIM. The underlines show the best performances among the results using the same joint filter. The bolds shows the best performance among all results.

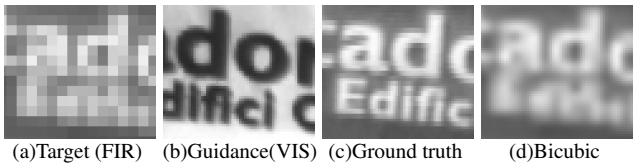
Method	$\sigma = 25$		$\sigma = 50$		$\sigma = 100$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
target	20.47	0.483	14.96	0.255	10.35	0.116
BM3D[10]	**31.88**	0.883	27.71	0.795	21.81	0.677
JBF[48]	27.72	0.763	25.48	0.717	21.55	0.676
DASC[31]+JBF[48]	28.61	0.806	26.64	0.766	21.69	0.633
Proposed (JBF[48])	31.40	0.880	27.03	0.771	21.82	0.685
MSJF[55]	25.84	0.715	24.91	0.704	21.31	0.671
DASC[31]+MSJF[55]	28.34	0.816	26.25	0.759	21.42	0.650
Proposed(MSJF[55])	28.72	0.819	26.55	0.775	21.67	0.688
GF[22]	27.33	0.766	25.47	0.737	21.41	0.699
DASC[31]+GF[22]	29.56	0.848	26.78	0.788	21.57	0.686
Proposed(GF[22])	31.44	**0.895**	**28.00**	**0.857**	**22.27**	**0.783**

adding Gaussian white noise (σ=25, 50 and 100) were used as the target image, whereas the clear NIR images were used as the guidance image.

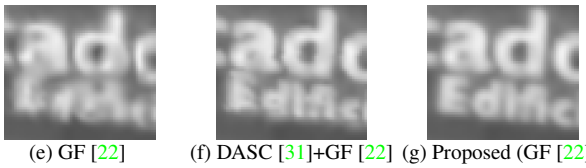
Comparison of the proposed method with the existing methods including BM3D [10] and GF [22] are shown in Fig. 6. To evaluate the performance of the existing joint filters with the cross-modal registration, we also used the aligned guidance image by applying DASC [31], which is a state-of-the-art registration method for cross-modal images pairs. Since DASC [31] is sensitive to noise, the noisy target images are denoised by using BM3D [10] before applying the DASC [31].

Figure 6 (d) shows that BM3D [10] cannot recover the texture on the statue. The result by GF [22] (Fig. 6 (e)) show that the blur artifacts are generated due to the misalignment. As shown in Fig. 6 (f), the existing joint filters [22] using the aligned guidance image also generates artifacts due to the alignment errors. On the other hand, Fig. 6 (g) show that the proposed method with these joint filters can recover the texture using the guidance image without artifacts. Note that the results by JBF [48] and MSJF [55] are presented in our supplemental materials due to the space limitations.

To evaluate the performance of the proposed method



(a)Target (FIR) (b)Guidance(VIS) (c)Ground truth (d)Bicubic



(e) GF [22] (f) DASC [31]+GF [22] (g) Proposed (GF [22])

Figure 7. FIR image up-sampling guided by visible image. ($\times 4$)

Table 2. Mean of RMSE and SSIM. The underlines show the best performances among the results using the same joint filter. The bolds show the best performance among all results.

Method	$\times 4$		$\times 8$	
	RMSE	SSIM	RMSE	SSIM
bicubic	3.981	0.9376	6.474	0.8836
JBF[48]	4.037	0.9387	6.587	0.8837
DASC [31]+JBF[48]	4.030	0.9389	6.599	0.8830
Proposed (JBF[48])	3.883	0.9396	6.273	0.8865
MSJF [55]	3.957	0.9397	6.600	0.8816
DASC [31]+MSJF [55]	3.990	0.9388	6.616	0.8820
Proposed (MSJF [55])	3.875	0.9398	6.266	0.8869
GF [22]	3.964	0.9391	6.644	0.8832
DASC [31]+GF [22]	3.979	0.9391	6.671	0.8823
Proposed (GF [22])	**3.874**	**0.9399**	**6.252**	**0.8873**

with the existing methods quantitatively, we measured PSNR and SSIM as shown in Table 1. These quantitative results show that 1) the performances are improved from the naive joint filters by applying the proposed method, and 2) the proposed method with GF [22] outperforms the compared methods.

4.2. Image up-sampling

Our framework is also effective for image up-sampling. We have already shown one result in Fig. 1. To evaluate the performance for the image up-sampling throughly, we used 100 image pairs of the visible gray images and FIR images collected by Aguilera*et al*. [1, 2]. We set the magnification rate as four and eight. The residual interpolation [30] was used as the post-processing.

An example of image up-sampling results is presented in Fig. 7. The result by the bicubic interpolation (Fig. 7 (d)) shows the limited visibility of the characters, *e.g*. "E", "d" and "f". The results by GF [22] without alignment (Fig. 7 (e)) show that the visibilities of the characters are degraded due to the misalignment. As shown in Fig. 7 (f), although the visibilities of the characters are improved by DASC [31], the discontinuity artifacts due to the subtle alignment error are generated on the character pattens. On the other hand, Fig. 7 (g) show that the proposed method can improve the visibilities while reducing the artifacts. Note that the results by JBF [48] and MSJF [55] are presented in our supplemental materials due to the space limitations.

To evaluate the performance of the proposed method and the existing method, root-mean-square error (RMSE) and SSIM were measured. As shown in Table 2, similar to the results on the image denoising in Sec. 4.1, the proposed method outperforms the existing methods, *i.e*. the joint filters [22, 55, 48] with/without DASC [31]. Note that, the proposed method with GF [22] shows the best performance among the all results.

4.3. Other applications

Finally, we demonstrate other applications of the cross-modal joint filter, where the input image pairs usually contain the misalignment.

Flash/no-flash photography: In the flash/no-flash photography, the misalignment often exists among flash and no-flash images [48]. The proposed method can generate the clear image from these misaligned flash/no-flash image pairs. An example of the results is shown in Fig. 8, where the flash/no-flash image pairs were captured by a mobile phone camera (GALAXY-SⅢ). As shown in Fig. 8 (a) and (b), the target no-flash image contains large noise, whereas the flashed guidance image is clear by the flash. Figure 8 (c) shows that the result by GF [22] is blurred due to the misalignment. On the other hand, Fig. 8 (d) shows that the proposed method can preserve the texture on the zebra doll while reducing the noise in the target image using the misaligned guidance image.

Haze removal: The proposed method is also applicable to haze removal by transferring the NIR image textures into the visible color image. We demonstrate the effectiveness of the proposed method using visible and NIR image pairs collected by Brown *et al*. [4]. The target visible color image, the guidance NIR image, the results by GF [22] and the proposed method with the GF [22] are shown in Fig. 9. The GF [22] generates the ghost artifacts on the mountain surface and at the boundary between the sky and the mountain region due to the misalignment. On the other hand, the proposed method can remove the haze effectively while reducing the ghost artifacts as shown in Fig. 9 (d).

Depth map interpolation: The proposed method can interpolate the depth map effectively without the accurate calibration and the temporal synchronization. We demonstrate the performance for the depth map interpolation using KITTI dataset which includes the visible color and the depth map image pairs [19]. Here, the depth maps contain the missing pixels. There are the misalignments between the visible color and the depth map image pairs because the captured time is slightly different. An example of the interpolated results by the proposed method is shown in Fig. 10. Here, we interpolate the dense depth map from sparse depth data [22]. Figure 10 (g) shows that GF [22] generates the blurry result with the ghost artifacts due to the misalignment on the fence region. Contrary to the naive use of the GF [22], the proposed method can interpolate the depth map without the artifacts as shown in Fig. 10 (h).

(a)Target(no-flash) (b)Guidance(flash) (c) GF [22] (d)Proposed(GF[22])

Figure 8. Flash/no-flash photography .

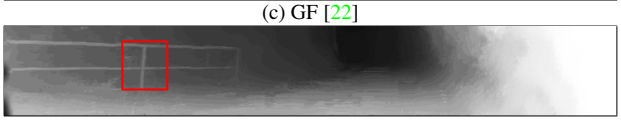
(a) Target (VIS) (b)Guidance(NIR) (c) GF [22] (d)Proposed(GF[22])

Figure 9. Results for haze removal using visible and NIR images.

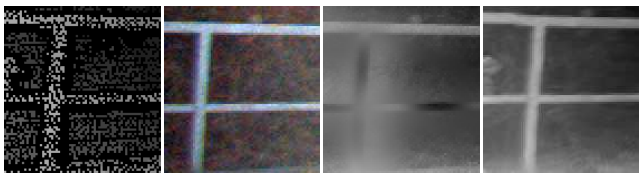
(a) Target (depth)

(b) Guidance (visible color)

(c) GF [22]



(d) Proposed (GF [22])



(e) Target (f) Guidance (g) GF [22] (h)Proposed(GF[22])

Figure 10. Depth map interpolation guided by visible image.

5. Conclusion

We have proposed the novel misalignment-robust joint filter. The proposed method can extend the existing cross-modal joint filters because the cross-modal cost volume are basically generated from the cost function of the existing cross-modal joint filters. The output image is composed by fusing the set of the filtered images with the weight volume. Experimental results have shown that the proposed method is effective for various applications such as image denosing, up-sampling, haze removal and depth map interpolation.

References

[1] C. A. Aguilera, F. J. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo. Learning cross-spectral similarity measures with deep convolutional neural networks. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016. 7

[2] C. A. Aguilera, A. D. Sappa, and R. Toledo. Lghd: A feature descriptor for matching across non-linear intensity variations. *In Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, 2015. 7

[3] M. Bleyer, C. Rhemann, and C. Rother. Patchmatch stereo-stereo matching with slanted support windows. *In Proc. of British Machine Vision Conference (BMVC)*, 11, 2011. 2

[4] M. Brown and S. Süsstrunk. Multi-spectral sift for scene category recognition. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1, 6, 8

[5] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2, 2005. 2

[6] S. C. Z. Feng, X. Zhang, L. Shen, and S. Süsstrunk. Near-infrared guided color image dehazing. *In Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, 2013. 1

[7] A. Chakrabarti and T. Zickler. Statistics of real-world hyperspectral images. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1

[8] Y. Choi, N. Kim, S. Hwang, and I. S. Kweon. Thermal image enhancement using convolutional neural network. *In Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2016. 1

[9] P. L. Combettes and J. C. Pesquet. Proximal splitting methods in signal processing. *Fixed-point algorithms for inverse problems in science and engineering*, 2011. 4

[10] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance-chrominance space. *In Proc. of Int. Conf. on Image Processing (ICIP)*, 2007. 6, 7

[11] L. Dai, M. Yuan, F. Zhang, and X. Zhang. Fully connected guided image filtering. *In Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2015. 2

[12] R. H. C. De Souza, M. Shimizu, M. Okutomi, and S. Yoshimura. Nonrigid registration based on projected joint entropy combined with gradient similarity. *Optical Engineering*, 49(12), 2010. 2

[13] M. Elad. On the origin of the bilateral filter and ways to improve it. *IEEE Trans. on Image Processing (TIP)*, 11(10), 2002. 5

[14] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Trans. on Graphics (TOG)*, 27(3):67, 2008. 2

[15] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. Image guided depth upsampling using anisotropic total generalized variation. *In Proc. of IEEE Int. Conf. on Computer Vision(ICCV)*, 2013. 1

[16] G. D. Finlayson and A. E. Hayes. Pop image fusion-derivative domain image fusion without reintegration. *In Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2015. 2

[17] C. Fredembach, N. Barbuscia, and S. Süsstrunk. Combining visible and near-infrared images for realistic skin smoothing. *Color and Imaging Conference*, (1), 2009. 1

[18] C. Fredembach and S. Süsstrunk. Colouring the near infrared. *In Proc. of the IS&T/SID 16th Color Imaging Conference*, 2008. 1

[19] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 8

[20] B. Ham, M. Cho, and J. Ponce. Robust image filtering using joint static and dynamic guidance. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[21] B. Ham, D. Min, and K. Sohn. Depth superresolution by transduction. *IEEE Trans. on Image Processing (TIP)*, 24(5), 2015. 1

[22] K. He, J. Sun, and X. Tang. Guided image filtering. *In Proc. of European Conf. on Computer Vision (ECCV)*, 2010. 1, 2, 4, 5, 6, 7, 8

[23] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(12), 2011. 2

[24] Y. S. Heo, K. M. Lee, and S. U. Lee. Robust stereo matching using adaptive normalized cross-correlation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(4), 2011. 2

[25] D. Herrera, J. Kannala, and J. Heikkilä. Joint depth and color camera calibration with distortion correction. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(10), 2012. 1

[26] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3), 1981. 2, 3

[27] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2), 2013. 2, 3

[28] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[29] M. Irani and P. Anandan. Robust multi-sensor image alignment. *In Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 1998. 2, 4

[30] D. Kiku, Y. Monno, M. Tanaka, and M. Okutomi. Residual interpolation for color image demosaicking. *In Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, 2013. 7

[31] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 3, 4, 5, 6, 7, 8

[32] S. Kim, D. Min, S. Lin, and K. Sohn. Deep self-correlation descriptor for dense cross-modal correspondence. *In Proc. of European Conf. on Computer Vision (ECCV)*, 2016. 2

[33] S. J. Kim, F. Deng, and M. S. Brown. Visual enhancement of old documents with hyperspectral imaging. *Pattern Recognition*, 44(7), 2011. 1

[34] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele. Joint bilateral upsampling. *ACM Trans. on Graphics (TOG)*, 26(3):96, 2007. 2

[35] D. Krishnan and R. Fergus. Dark flash photography. *ACM Trans. on Graphics (TOG)*, 28(3)(96), 2009. 5

[36] D. Krishnan and R. Fergus. Fast image deconvolution using hyper-laplacian priors. *In Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2009. 4

[37] H. Kwon and Y.-W. Tai. Rgb-guided hyperspectral image upsampling. *In Proc. of IEEE Int. Conf. on Computer Vision(ICCV)*, 2015. 1

[38] S. Z. Li, R. F. Chu, S. C. Liao, and L. Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(4), 2007. 1

[39] X. Li. Demosaicing by successive approximation. *IEEE Trans. on Image Processing (TIP)*, 14(3), 2005. 6

[40] Y. Li, D. Min, M. N. Do, and J. Lu. Fast guided global interpolation for depth and motion. *In Proc. of European Conf. on Computer Vision (ECCV)*, 2016. 1

[41] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(5), 2011. 2, 3

[42] S. Lu, X. Ren, and F. Liu. Depth enhancement via low-rank matrix completion. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[43] Y. M. Lu, , C. Fredembach, M. Vetterli, and S. Süsstrunk. Designing color filter arrays for the joint capture of visible and near-infrared images. *In Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, 2009. 1

[44] P. Milanfar. A tour of modern image filtering: New insights and methods, both practical and theoretical. *IEEE Signal Processing Magazine*, 30(1), 2013. 5

[45] N. J. Morris, S. Avidan, W. Matusik, and H. Pfister. Statistics of infrared images. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007. 1

[46] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3), 2013. 4

[47] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. S. Kweon. High-quality depth map upsampling and completion for rgb-d cameras. *IEEE Trans. on Image Processing (TIP)*, 23(12), 2014. 1

[48] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama. Digital photography with flash and no-flash image pairs. *ACM Trans. on Graphics (TOG)*, 23(3), 2004. 1, 2, 4, 5, 6, 7, 8

[49] J. P. Pluim, J. A. Maintz, and M. A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Trans. on Medical Imaging*, 22(8), 2003. 2

[50] Z. Sadeghipoor, Y. M. Lu, and S. Süsstrunk. Correlation-based joint acquisition and demosaicing of visible and near-infrared images. *In Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, 2011. 1

[51] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision (IJCV)*, 47(1-3), 2002. 2, 3

[52] L. Schaul, C. Fredembach, and S. Süsstrunk. Color image dehazing using the near-infrared. *In Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, 2009. 1

[53] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007. 2

[54] X. Shen, L. Xu, Q. Zhang, and J. Jia. Multi-modal and multi-spectral registration for natural images. *In Proc. of European Conf. on Computer Vision (ECCV)*, 2014. 1, 2

[55] X. Shen, C. Zhou, L. Xu, and J. Jia. Mutual-structure for joint filtering. *In Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2015. 2, 4, 5, 6, 7, 8

[56] T. Shibata, M. Tanaka, and M. Okutomi. Unified image fusion based on application-adaptive importance measure. *In Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, 2015. 1

[57] T. Shibata, M. Tanaka, and M. Okutomi. Visible and near-infrared image fusion based on visually salient area selection. *In Proc. of SPIE Electrical Imaging*, 2015. 1

[58] T. Shibata, M. Tanaka, and M. Okutomi. Gradient-domain image reconstruction framework with intensity-range and base-structure constraints. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[59] S. Süsstrunk and C. Fredembach. Enhancing the visible with the invisible: Exploiting near-infrared to advance computational photography and computer vision. *SID Int. Symposium Digest*, 41(1), 2010. 1

[60] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Processing (TIP)*, 13(4), 2004. 6

[61] S. Wug Oh, M. S. Brown, M. Pollefeys, and S. Joo Kim. Do it yourself hyperspectral imaging with everyday digital cameras. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[62] Q. Xie, Q. Zhao, D. Meng, Z. Xu, S. Gu, W. Zuo, and L. Zhang. Multispectral images denoising by intrinsic tensor sparsity regularization. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[63] Q. Yan, X. Shen, L. Xu, S. Zhuo, X. Zhang, L. Shen, and J. Jia. Cross-field joint image restoration via scale map. *In Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2013. 2, 5

[64] Q. Yang. A non-local cost aggregation method for stereo matching. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[65] L. Yun, R. Jeffrey, and H. Dieter. Pedestrian detection in near-infrared night vision system. *IEEE Intelligent Vehicles Symposium (IVS)*, 2010. 1

[66] C. Zhang and Z. Zhang. Calibration between depth and color sensors for commodity depth cameras. *Computer Vision and Machine Learning with RGB-D Sensors*, 2014. 1

[67] M. M. Zhang, J. Choi, K. Daniilidis, M. T. Wolf, and C. Kanan. Vais: A dataset for recognizing maritime imagery in the visible and infrared spectrums. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015. 1

[68] Q. Zhang, X. Shen, L. Xu, and J. Jia. Rolling guidance filter. *In Proc. of European Conf. on Computer Vision (ECCV)*, 2014. 2

[69] X. Zhang, T. Sim, and X. Miao. Enhancing photographs with near infra-red images. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008. 1

[70] S. Zhuo, X. Zhang, X. Miao, and T. Sim. Enhancing low light images using near infrared flash images. *In Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, 2010. 1, 2