Peer reviewed version

Link to published version (if available):
10.1109/ICIP46576.2022.9897779

Link to publication record in Explore Bristol Research
PDF-document

University of Bristol - Explore Bristol Research
General rights

UNSUPERVISED IMAGE FUSION USING DEEP IMAGE PRIORS

Xudong Ma, Paul Hill, Nantheera Anantrasirichai, Alin Achim

Visual Information Laboratory, University of Bristol, Bristol, UK

ABSTRACT

A significant number of researchers have applied deep learning methods to image fusion. However, most works require a large amount of training data or depend on pre-trained models or frameworks to capture features from source images. This is inevitably hampered by a shortage of training data or a mismatch between the framework and the actual problem. Deep Image Prior (DIP) has been introduced to exploit convolutional neural networks' ability to synthesize the 'prior' in the input image. However, the original design of DIP is hard to be generalized to multi-image processing problems, particularly for image fusion. Therefore, we propose a new image fusion technique that extends DIP to fusion tasks formulated as inverse problems. Additionally, we apply a multi-channel approach to enhance DIP's effect further. The evaluation is conducted with several commonly used image fusion assessment metrics. The results are compared with state-of-the-art image fusion methods. Our method outperforms these techniques for a range of metrics. In particular, it is shown to provide the best objective results for most metrics when applied to medical images.

Index Terms— image fusion, unsupervised learning, inverse problem, deep image priors

1. INTRODUCTION

Image fusion is the process of combining information from multiple images into a single representation [1]. It is widely used in many application domains, such as medical imaging [2], photography [3], and surveillance [4]. A large amount of research has been done in this area. Conventionally, image fusion can be performed at pixel-level, feature-level, or decision-level [5]. Pixel-level image fusion has been extensively studied because of its high efficiency [5]. It can be performed using either fixed transform-based methods or via learnt Sparse Representations (SR) [5]. If using fixed transforms, the original images need to be decomposed into other domains using methods such as the Discrete Wavelet Transform (DWT) or the contourlet transform [6]. Decomposition coefficients are fused using fusion-rules such as weighted-averaging [7] and maximum selection [8]. Finally, the fused image is generated by taking the inverse transform of the fused coefficients. For SR-based methods, a dictionary is learnt before fusion. Researchers have proposed many ways to obtain the optimal dictionary. Most utilize either mathematical models or sample learning [9]. The input images are first decomposed and sparse coded during fusion. Different strategies have been developed to achieve this, for example, group SR [10] and gradient constrained SR [11]. Finally, the computed sparse coefficients are fused and reconstructed into a single image according to various fusion rules [12].

The fusion task can be seen as an inverse problem. Solving with mathematical models, various loss functions and optimization strategies have been proposed to get the best fusion results. For this kind of method, the loss function plays a vital role in the performance achieved. The L_1 norm is widely used as a regularization component of the loss function [13]. However, it has some inherent weakness that often causes the final result to be over-smoothed. To address this, a non-convex penalty regularization technique was proposed in [14] and achieved state-of-the-art image fusion performance. With the development of deep Convolutional Neural Networks (CNNs), researchers have started using such techniques for image fusion. Liu *et al* [15] proposed the application of blurred image patches to train a decision map using CNNs. The fused image was generated based on the decision map and the input images. Li *et al* [16] enhanced Liu's method by introducing a multi-layer feature extraction technique. However, like most other deep learning based fusion models, they all depend on training data or pre-training. Nevertheless, training data is hard to obtain for image fusion tasks. Furthermore, it is also challenging to find well-fitted pre-trained models for some specific applications, such as the fusion of medical images. Most models are pre-trained on natural images, which are very different from medical images. This inherently limits performance in particular applications.

Unsupervised learning is a way to address the issues of current deep learning based fusion methods. This paper proposes a new technique, based on Deep Image Prior (DIP) [17], to fuse images in an unsupervised way. DIP assumes that a significant number of underlying image statistical priors can be obtained by the structure of CNNs even without any training [17]. The authors employed an encoder-decoder network to verify their assumption on several single image restoration tasks. Although they also provided an example of applying DIP to multi-image processing problems, their design is still hard to be extended to problems such as image fusion.

Here, we pose image fusion as an inverse problem and design a loss computed from the output and all source images directly. This makes it easy to tackle multi-image processing problems. In addition, this paper also proposes a multi-channel training strategy to enhance the image prior extraction ability of DIP. Finally, the performance of our approach is compared with the non-convex penalty based fusion method in [14] and two deep learning based methods [16] [18], which are recently published state-of-the-art model-based and data-driven image fusion techniques, respectively. Better results are obtained visually and objectively for the fusion of both mono-modal images and cross-modal images.

2. METHODOLOGY

2.1. Inverse Problem Formulation

Typically, image fusion is viewed as a problem of generating the fused image from several source images. However, the fusion task can also be posed as an inverse problem. It assumes that the fused image X_0 exists already, and the image formation model describes how to obtain the source images X_i from the fused image as $X_i = f(X_0)$, where f is related to some degradation operations. When the source images are viewed as signals coming from different sensors, the problem can be modeled as [19].

$$X_i = \beta_i X_0 + N_i \qquad (1)$$

where β_i and N_i are the sensor gain and sensor noise of the ith source image. X_0 is the fused result. In this paper, the sensor gain β_i is estimated from the source images X_i employing a patch-based principal component analysis (PCA) strategy. The effectiveness and robustness of PCA on computing sensor gains have been proven over the years [19, 20]. The gains β_i (see Fig. 1) constitute the forward operators that transform X_0 into the source images. The goal, therefore, becomes to minimize the differences between each source image and the degraded fused image, as shown in (2).

$$L = \|X_1 - \beta_1 X_0\|_2^2 + \|X_2 - \beta_2 X_0\|_2^2 \qquad (2)$$
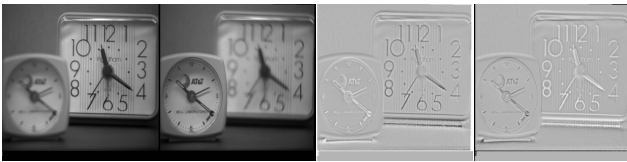
Fig. 1: Source image and sensor gain examples. The first two images are the sources. The third and fourth are their corresponding gains computed by PCA method.

2.2. DIP-based Image Fusion

DIP is a classical encoder-decoder style network where only one image is directly employed to compute the loss. Therefore, it is always applied in single image processing problems. Although Ulyanov *et al* [17] have provided an example of using DIP to reconstruct an image from a pair of flash and no-flash images, their method is not suitable for most image fusion tasks where we want to harvest information from all the source images optimally. In their method, the flash image is applied as the network input, and only the no-flash one is employed to compute the loss. Therefore, their strategy is not straightforwardly extended to multi-image problems.

This paper proposes to directly compute the loss from all source images. As shown in Fig. 2, the input of the network, I, is a random tensor that has the same shape as the source images X_1 and X_2. After passing the input through the DIP architecture (expressed as the weights matrix W), the output X_0 is generated (3).

$$X_0 = WI \qquad (3)$$

When formulating the image fusion task as an inverse problem discussed in the previous section, the loss of the network can be computed from (2). Subsequently, the backpropagation updates W, and a new iteration begins. After a judiciously chosen number of iterations, the output with minimum loss value will be selected as the fused image.

Our loss function offers DIP a chance to harvest information meaningfully from all source images. Furthermore, this structure is easily generalized to fusing images with more than two sources. Additionally, This method trains from scratch for each set of source images and only harvests the underlying features of the current sources. This inherently addresses the lack of training data in image fusion, and therefore overfitting is no longer an issue.
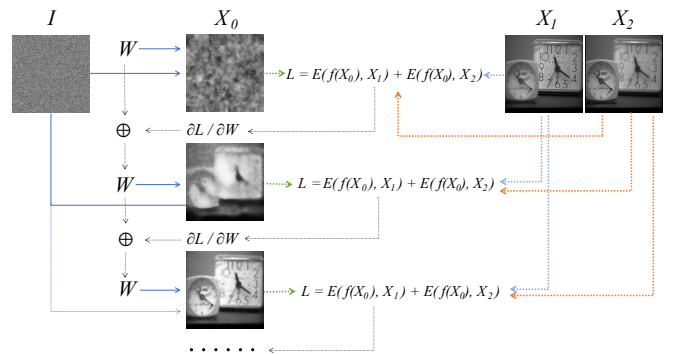
Fig. 2: The image fusion process. I is the input of the network. W and X_0 are the parameter and output of the network, respectively. X_1 and X_2 are the source images. $E()$ is the loss function corresponding to each input. $f()$ is the degradation function and L is the overall loss.

2.3. Network Architecture

This paper employs the super-resolution structure of DIP from [17] as the foundation of the network. The hyperparameters are kept as their default values. Both the encoder and decoder include 5 convolutional layers of 128 dimensions. For each skip connection, only 4 dimensions are left as the aid. The learning rate is set to 0.01. LeakyReLU and Adam are applied as the activation function and optimizer, respectively. The input and initialization of the network are done randomly. We attempted to perform the task on deeper networks with higher dimensions for each layer, and we also tried different learning rates, activation functions, optimizers. However, they do not help improve the results significantly.

2.4. Multi-Channel Training Strategy

The original DIP network only provides a single channel output. In order to capture image priors as exhaustively as possible, we apply a multi-channel strategy to train the network. When the total number of output channels is set to n, we copy the source images into n channels. The input becomes a n-channel random tensor, and the output is adjusted to n channels as well. The loss function is still calculated by (2) for each output channel. Subsequently, an averaging strategy is employed to get the final result X_0', as shown in (4), where X_0^i is the ith channel of the output.

$$X_0' = \frac{\sum_{i=1}^{n} X_0^i}{n} \tag{4}$$

This alternative is an efficient way to train the network several times and combine the results to achieve better performance. As shown in Fig. 3, our multi-channel strategy outperforms the single channel method of the original DIP network.

3. EXPERIMENTS AND EVALUATION

3.1. Experiments

The experiments rely on 12 greyscale image pairs of different categories: 1 multi-focus image pair, 1 pair of infrared and visible images, 9 pairs of Magnetic Resonance (MR) and Computed Tomography (CT) from different patients, and a pair of MR and Ultrasound (US) phantom images. The MR and US image pair was obtained through the combination of a beefsteak and a polyvinyl alcohol (PVA) phantom [21]. The results are compared with three state-of-the-art image fusion methods, one model-based strategy and two deep learning based methods. In the model-based one [14], the authors formulated the problem as an inverse one and applied nonconvex penalties to get better results than many state-of-the-art methods. The second benchmark [16] decomposed source images into base and detail parts (Base and Detail Net). They applied a VGG net trained on ImageNet to harvest deep features for the detail parts and fused the features finally. Our last

benchmark is DenseFuse [18]. They added a dense block into a pre-trained encoder to capture deep features exhaustively before fusing the features using some strategies. The second benchmark can only process 256×256 images, so we resize all source images. The number of training iterations is set to 2000, an empirical number obtained by trial and error.
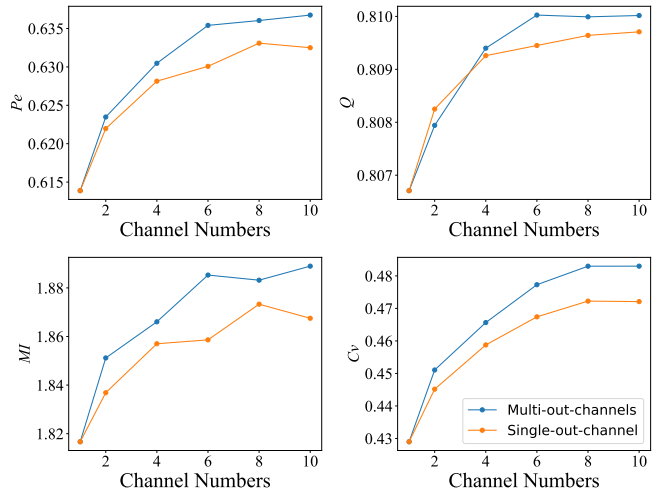

Fig. 3: The average objective results, on the 12 experiment image pairs, of our multi-channel training strategy and the single-channel method of the original DIP network.

3.2. Objective and Visual Evaluation

This paper employs four commonly used numerical measures for image fusion to evaluate the performance objectively: Petrovic and Xydeas's metric (Pe) [22], Mutual Information (MI) [23], Piella's metric (Q) [24] and Cvejic's Quality Index (Cv) [25]. Pe quantifies how much important visual information from the source images is transferred in the fused image. Q is a metric considering local information and the level of distortion. MI focuses on the shared information present in both the fused and the source images. Cv is based on a universal image quality index, and it uses local measurements to quantify the importance of the information. None of them is suitable for all images and can assess the fusion performance comprehensively. Therefore, this paper employs all of them to make the evaluation more comprehensive.

The results corresponding to training using the different number of channels are shown in Fig. 3. The results on all image pairs are averaged for each channel number. It is clear that our multi-channel strategy performs better on all metrics than the single channel method of the original DIP. This ascertains that our method can better capture the underlying image statistics than the original DIP. Also, it objectively proves the efficiency of our DIP-based multi-channel training strategy.

We illustrate the performance of different methods in table 1. The results of 9 pairs of CT and MR images are averaged to save space. It is clear that our 10-channel training strategy
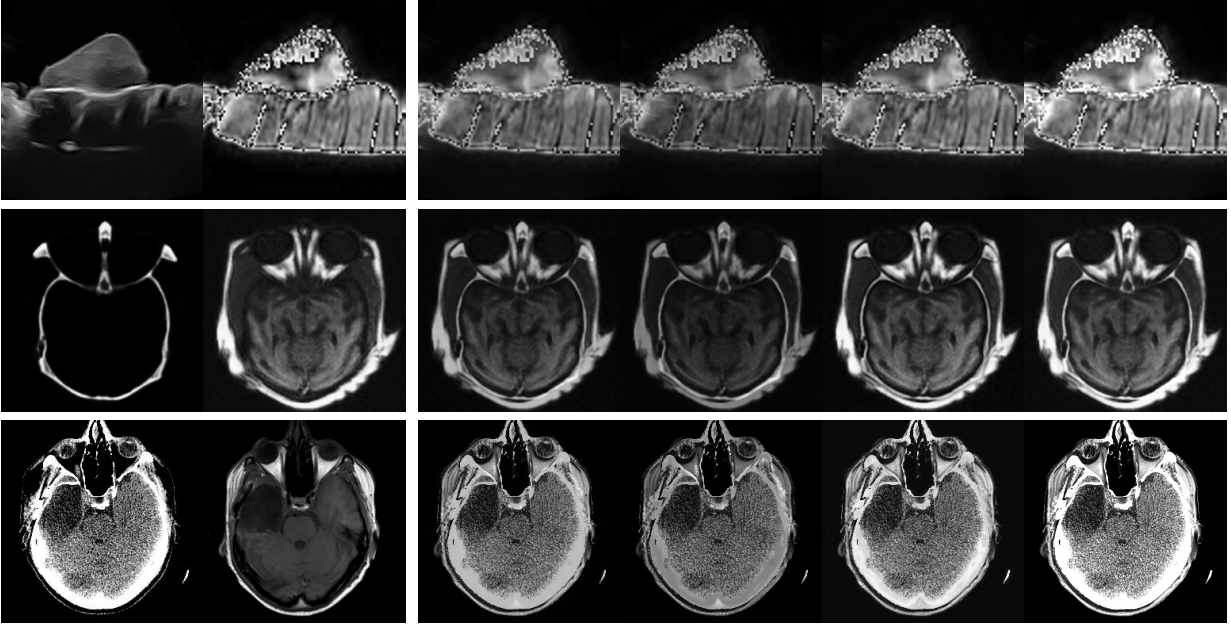
Fig. 4: The results of different fusion methods. From left to right, two source images and the results for the non-convex penalty based strategy, the Base and Detail Net, the DenseFuse Net and our 10-channel training DIP strategy, respectively. The Source image in the first row is a pair of MR and US phantom images obtained through the combination of a beefsteak and a polyvinyl alcohol (PVA) phantom. The other two rows are CT and MR image pairs of the human head.

Table 1: The objective results. Average operation is carried out to the CT and MR image pairs to save space. The highest scores are bold and the second-highest scores are shown in blue.

		Non-Convex Penalty	Base and Detail Net	DenseFuse	Our 10-channel Method
Multi focus	Pe	0.6297	**0.6483**	0.6291	0.6375
	MI	**3.1107**	3.0852	3.0052	3.063
	Q	0.8843	**0.9049**	0.8916	0.8487
	Cv	0.7278	**0.7909**	0.7779	0.6991
VIS and IR	Pe	0.4606	**0.4639**	0.4338	0.4464
	MI	1.4172	1.2546	**1.4817**	1.3017
	Q	0.6932	**0.7279**	0.6657	0.6658
	Cv	0.6941	**0.7085**	0.6126	0.6785
Phantom Data	Pe	0.5468	0.4957	0.5934	**0.6625**
	MI	1.9511	1.6075	1.8493	**1.9533**
	Q	0.8078	0.7932	0.831	**0.8474**
	Cv	0.5555	0.6005	0.5476	**0.6152**
CT MR Average	Pe	0.5349	0.4383	0.5723	**0.6548**
	MI	**1.9769**	1.6052	1.6962	1.8177
	Q	0.7761	0.7263	0.7994	**0.8176**
	Cv	**0.4325**	0.3798	0.3664	0.4223

achieves the highest or second-highest scores in more than half of the cases. Moreover, Our approach shows apparent advantages compared to others when fusing medical images. For example, we improve the value of Pe by about 15% over the second-best method on both the phantom and real medical images. Especially on the noisy phantom data, our approach shows the best results on all metrics. Although the benchmarks outperform us on several assessments for the other two image pairs, our results are trailing very close behind.

As shown in Fig. 4, visually, our fused images (the 6th column) can capture the bright information from the source

images better and show better contrast than the benchmarks. When the source images are extremely noisy, such as the phantom data (the 1st row), this strength is more apparent.

4. CONCLUSIONS

This paper introduces an unsupervised method for image fusion. We pose the task as an inverse problem and enable DIP to be used in multi-sensor image fusion problems. Our method implicitly solves the problem of data scarcity in previous deep learning based methods. Furthermore, as all training is conducted specifically for the source images at hand, overfitting and model mismatch problems in existing deep learning based image fusion methods are eliminated. Additionally, the use of the multi-channel training strategy further strengthens the ability of DIP to capture image statistical priors. As a result, our approach is more robust to noise than its competitors. Visually, we provide fusion results that are clearer and have better contrast than the benchmarking state-of-the-art conventional and deep learning based methods. Objectively, the method proposed in this paper achieves the best performance in terms of several metrics. The advantages are particularly evident when fusing medical images especially when noisy. Finally, Our approach can be easily extended to fuse image datasets containing more than two images, and our multi-channel training strategy is also applicable to other deep learning based problems.

5. REFERENCES

[1] P. Hill, "Image fusion using complex wavelets," *proc bmvc*, 2003.

[2] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Information Fusion*, vol. 19, pp. 4–19, 2014.

[3] C. Y. Wen and J. K. Chen, "Multi-resolution image fusion technique and its application to forensic science," *Forensic Science International*, vol. 140, no. 2-3, pp. 217–232, 2004.

[4] N. Paramanandham and K. Rajendiran, "Multi sensor image fusion for surveillance applications using hybrid image fusion algorithm," *Multimedia Tools and Applications*, vol. 77, no. 10, pp. 12405–12436, 2018.

[5] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Information Fusion*, vol. 33, 2017.

[6] S. Yang, W. Min, Z. W. Liu, and L. Jiao, "Image fusion using a nsdfb-based contourlet packet," in *International Symposium on Intelligent Signal Processing and Communication Systems*, 2008.

[7] J. H. Jang, Y. Bae, and J. B. Ra, "Contrast-enhanced fusion of multisensor images using subband-decomposed multiscale retinex," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3479–90, 2012.

[8] J. Hu and S. Li, "The multiscale directional bilateral filter and its application to multisensor image fusion," *Information Fusion*, vol. 13, no. 3, pp. 196–206, 2012.

[9] R. Rubinstein, S. Member, IEEE, A. M. Bruckstein, and Member, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[10] S Li, H Yin, and L Fang, "Group-sparse representation with dictionary learning for medical image denoising and fusion," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 12, pp. 3450–3459, 2012.

[11] C. Chen, Y. Li, L. Wei, and J. Huang, "Image fusion with local spectral consistency and dynamic gradient sparsity," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[12] M. Kim, D. K. Han, and H. Ko, "Joint patch clustering-based dictionary learning for multimodal image fusion.," *Acm Transactions on Sensor Networks*, vol. 6, no. 3, pp. 20, 2010.

[13] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing*, Mathematical problems in image processing. Partial differential equations and the calculus of variations. Foreword by Olivier Faugeras. 2nd ed, 2002.

[14] N. Anantrasirichai, R. Zheng, I. Selesnick, and A. Achim, "Image fusion via sparse regularization with non-convex penalties," *Pattern Recognition Letters*, vol. 131, 2020.

[15] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.

[16] L. Hui, X. J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *International Conference on Pattern Recognition 2018*, 2018.

[17] D Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[18] L. Hui and X. J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[19] M. Kumar, "Total variation regularization-based pixel level image fusion," in *Signal Processing Systems*, 2010.

[20] Jin Chen, Jose Nunez-Yanez, and Alin Achim, "Joint video fusion and super resolution based on markov random fields," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 2150–2154.

[21] O. E. Mansouri, F. Vidal, A. Basarab, P. Payoux, and J. Y. Tourneret, "Fusion of magnetic resonance and ultrasound images for endometriosis detection," *IEEE Transactions on Image Processing*, vol. PP, no. 99, pp. 1–1, 2020.

[22] C. S. Xydeas and V. Pv, "Objective image fusion performance measure," *Military Technical Courier*, vol. 56, no. 4, pp. 181–193, 2000.

[23] J. P. W. Pluim and J. B. A. Maintz, "Image registration by maximization of combined mutual information and gradient information," *IEEE Transactions on Medical Imaging*, vol. 19, no. 8, pp. P.809–814, 2000.

[24] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *International Conference on Image Processing*, 2003.

[25] N. Cvejic, A. Loza, D. R. Bull, and C. N. Canagarajah, "A novel metric for performance evaluation of image fusion algorithms," in *International Enformatika Conference*, 2005.