Pattern Recognition 88 (2019) 431–446

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Convolutional network architectures for super-resolution/sub-pixel mapping of drone-derived images

Pattathal V. Arun[a], Ittai Herrmann[b], Krishna M. Budhiraju[a], Arnon Karnieli[c,*]

[a] Indian Institute of Technology Bombay, India
[b] Hebrew University of Jerusalem, Israel
[c] Ben-Gurion University, Israel

ARTICLE INFO

Article history:
Received 19 June 2017
Revised 22 November 2018
Accepted 27 November 2018
Available online 28 November 2018

Keywords:
Sub-pixel mapping
Super-resolution
Convolutional neural network
Class distribution
Drone
UAV

ABSTRACT

Spatial resolution enhancement is a pre-requisite for integrating unmanned aerial vehicle (UAV) datasets with the data from other sources. However, the mobility of UAV platforms, along with radiometric and atmospheric distortions, makes the task difficult. In this paper, various convolutional neural network (CNN) architectures are explored for resolving the issues related to sub-pixel classification and super-resolution of drone-derived datasets. The main contributions of this work are: 1) network-inversion based architectures for super-resolution and sub-pixel mapping of drone-derived images taking into account their spectral-spatial characteristics and the distortions prevalent in them 2) a feature-guided transformation for regularizing the inversion problem 3) loss functions for improving the spectral fidelity and inter-label compatibility of coarser to finer-scale mapping 4) use of multi-size kernel units for avoiding over-fitting. The proposed approach is the first of its kind in using neural network inversion for super-resolution and sub-pixel mapping. Experiments indicate that the proposed super-resolution approach gives better results in comparison with the sparse-code based approaches which generally result in corrupted dictionaries and sparse codes for multispectral aerial images. Also, the proposed use of neural network inversion, for projecting spatial affinities to sub-pixel maps, facilitates the consideration of coarser-scale texture and color information in modeling the finer-scale spatial-correlation. The simultaneous consideration of spectral bands, as proposed in this study, gives better super-resolution results when compared to the individual band enhancements. The proposed use of different data-augmentation strategies, for emulating the distortions, improves the generalization capability of the framework. Sensitivity of the proposed super-resolution and sub-pixel mapping frameworks with regard to the network parameters is thoroughly analyzed. The experiments over various standard datasets as well as those collected from known locations indicate that the proposed frameworks perform better when compared to the prominent published approaches.

1. Introduction

Spatial resolution enhancement of remote sensing images has enjoyed the wide attention of researchers and is a pre-requisite for the integration and processing of datasets from diverse sources. Unlike other images, an increased spectral dimension (more than one spectral band), along with the geometric and atmospheric effects, present additional constraints for the resolution enhancement of unmanned aerial vehicle (UAV) datasets. Among the two different types of approaches relevant in this context, the super-resolution techniques attempt the reconstruction of finer-scale images from coarser ones, whereas the sub-pixel mapping approaches deal with the prediction of finer-scale fractional classified maps.

Most of the sub-pixel mapping methods [42,45,46,56,58,65] predict the target scale distributions merely based on fractional abundances, ignoring the coarser-level pixel distributions. Prominent single image super-resolution methods [5,13,24,39,61,64] approach the problem as an image transformation task where a feed-forward convolutional network is trained to learn the mapping between low-resolution and high-resolution image pairs. Although the effectiveness of CNN for image reconstruction is well established, the proper choice of network architecture for resolution enhancement of multi-spectral aerial images is still ambiguous.

While deep learning approaches seem to be computationally complex, once trained, the execution times of these methods are

* Corresponding author.
E-mail address: karnieli@bgu.ac.il (A. Karnieli).

https://doi.org/10.1016/j.patcog.2018.11.033
0031-3203/

comparable to the other approaches. Moreover, the significant improvement in accuracy as well as the effectiveness in modeling aerial-image features and handling distortions and artifacts, motivate the use of deep convolutional architectures for processing drone-derived images. In this regard, we explore novel architectures for offline super-resolution and sub-pixel classification of UAV datasets. In contrast to the recent CNN based strategies, the network-inversion techniques [16,41,49,50] are employed to model the spatial-spectral distributions across spatial-scales. This approach is found to be faster, and better trainable, when compared to the other prominent deep-learning based techniques. In addition, it avoids the bottleneck, due to increased input spectral-dimension, prevalent in the sparse-coding-based approaches. This study also explores the use of data augmentation, for emulating the distortions and variations of drone-derived datasets, for improving the generalization capability of the proposed frameworks.

2. Related work

This section reviews the prominent work related to the proposed sub-pixel classification and super-resolution strategies. In addition to the recent super-resolution and sub-pixel mapping approaches, review of the recent convolutional architectures, loss functions, and neural network inversion approaches is also presented. Based on the literature, the specific contributions of this research and novelty of the framework with respect to the existing approaches are discussed in Section 2.5.

2.1. Sub-pixel mapping

Most of the recent sub-pixel mapping techniques [42,46,58,59,65,77] attempt to optimize the spatial contiguity of classes without considering their distributions at coarser scales. Unlike these conventional strategies, geostatistical methods use the correlation between finer and coarser scale variograms to predict the sub-pixel labels [6,30,62]. However, variogram based approaches have poor generalization capability and mostly cause artifacts at higher scale factors. In this regard, several machine learning strategies such as genetic algorithm [57], Hopfield networks [56,63], feedforward neural networks [45] and CNNs [7,10,12,39] have been employed. Among these, the CNN based approaches [5,7,12,39] have reported the state-of-the-art results. In their work, Arun et al. [5] used a pre-trained CNN for projecting the fractional approximations (derived from the coarse image) to finer-scale classified maps. In a more recent work, Arun et al. [7] employed CNN for simultaneous optimization of spectral unmixing and mapping stages. Although these approaches yield satisfactory results, overfitting of the network affects the outcomes, which deteriorate further with the increase in scene complexity and scale factor. Also, the inter-band misalignments and other distortions affect the applicability of prominent sub-pixel mapping methods on drone-derived images.

2.2. Super-resolution

Classical single image super-resolution techniques often cause aliasing artifacts resulting in blurry outputs. Recently, various approaches, such as those based on local parametric regression [11,23,29,36], shape and texture modeling [1], recurrent patch learning [21,24,37,52,64], dictionary-based modeling [9,47,66,67,71,72], and neural networks [5,6,12,34,38,39], have been extensively explored for resolving these issues. Among these approaches, the CNN-based ones [12,34,38,39,61] illustrate the capability of CNNs in encoding the sparse-coding framework in an optimized manner. Nowadays, deeper networks are being explored for improving both reconstruction accuracy and performance. Liebel et al. [39] extended the work of Dong et al. [12] for remote sensing images. However, most of these techniques cannot be extended to the multispectral domain due to the increased computational cost, and also because they ignore the simultaneous learning of spectral bands [25,27,28,40,73]. In the context of hyperspectral image super-resolution, Dong et al. [13] yield satisfactory results. Relatively fewer attempts are seen on the super-resolution of drone-derived images, particularly due to the multi-spectral nature of data as well as the geometric and radiometric distortions due to the mobility of UAV platforms.

2.3. CNN architecture and loss functions

Following the success of CNNs, many researchers [20,26,33,51,53,54] have discussed the modeling of available network architectures for different image transformation tasks. He et al. [20] suggested the use of residual networks in which the additive merging of signals is explored. Similarly, convolutional inception architectures [53–55] allow the use of multi-size kernels in the same convolution units. Recently, Szegedy et al. [54] proposed an improvement on the inception nets, where residual connections are introduced between the inception units. However, these advancements in residual and inception networks have not been explored for super-resolution/sub-pixel classification.

While most of the CNN-based single image super-resolution approaches use root mean squared error (RMSE) based loss functions, Sajjadi et al. [48] illustrated that these functions generally result in artefacts as they compute the mean of many possible solutions. Recent studies [14,32] suggest that the shifting of loss computation from the image space to higher-level feature space results in sharper reconstructions [32]. The use of generative adversarial networks [19], in which the generator network simultaneously learns to fool a discriminator, has been thoroughly explored for various image analyses [19,35,54,60,74]. Ledig et al. [35] illustrated that the use of adversarial loss, in addition to the per-pixel losses, improves the reconstruction. The current study explores prominent loss functions for improving the proposed frameworks.

2.4. Neural network inversion

The neural network inversion techniques [3,31,44,75] are generally employed for the reconstruction of original inputs from corresponding network outputs. Recently, Dosovitskiy and Brox [15] studied the various image representations by inverting them with an up-convolutional neural network. Among the various features, the histogram of oriented gradients provided the best reconstruction. The authors also illustrated that the colors and rough contours of an image can be reconstructed even from higher layer activations. Xu et al. [69] proposed a deep learning architecture to capture the characteristics of image degradation. Similar applications of network inversion have been explored in [16,41,49,50]. Inspired from these approaches, this study investigates the applicability of network inversion strategies for resolving the issues prevalent in sub-pixel mapping and super-resolution of aerial images.

2.5. Contribution

The current study focuses on CNN-based offline sub-pixel classification and super-resolution, specifically for drone-derived images. We illustrate that the conventional CNN architectures cannot scale well for the enhancement of multispectral drone-derived images. In this regard, improved convolutional frameworks are proposed for sub-pixel classification and super-resolution. As far as we know, use of network inversion was not previously published for these tasks. The basic hypothesis here is that since the coarse-scale images can be regarded as convolved versions of the finer-

scale ones, the inversion technique can be utilized to learn the reconstruction of the latter, as well as the finer-scale fractional maps, from the former. The proposed frameworks work well for the cases even when the sparse-coding approaches fail to generalize the dictionaries and sparse codes. Also, the inversion technique avoids the bottleneck prevalent in sparse-coding strategies, and scales well for multispectral as well as hyperspectral images. Different atmospheric and geometric distortions are taken into account by augmenting the training datasets with different transformations to emulate these distortions. The applicability and advantages of inception and residual architectures, as well as network inversion, in overcoming the inter-band geometric misalignments are also investigated. Furthermore, the proposed super-resolution approach better preserves the spectral fidelity when compared to the existing strategies. Unlike in Wang et al. [61], in this work, an ensemble-based super-resolution strategy is simulated in a computationally optimal way by increasing the kernel diversity and extending the dropout concept for kernel selection. Additionally, an optimal initial upscaling strategy is proposed instead of the direct bi-cubic interpolation. This study also reviews the effect of available loss functions on both computational performance and accuracy of super-resolution and sub-pixel classification frameworks.

3. Approach

Consider a multispectral UAV image Y with a spatial-resolution (pixel size $s \times s$), and let W be its high-resolution counterpart with a pixel size $r \times r$ ($r \ll s$). The proposed sub-pixel mapping method aims to generate from Y, a fractional classified image I_k (for each class k) at a resolution r such that I_K is similar to the actual ground truth map (A_k) derived from W. It should be noted that the I_k denotes the classified map at a finer resolution in which only the kth class is labeled as 1, and all the other classes as zero. It is different from the coarser resolution fractional image (F_K) that gives the abundance fraction of the kth class at each pixel. Unlike the sub-pixel mapping method, the proposed super-resolution technique attempts to reconstruct a fine resolution image (W') from Y such that W' and W are similar. The RMSE, computed between the reconstructed fractional class image (I_K) and its corresponding ground truth (A_k), is used as the loss function for sub-pixel classification. Similarly, for super-resolution, the network attempts to minimize the inverse of average spectral similarity between the original high-resolution image (W) and its reconstructed version (W').

3.1. Dataset specifications

In this study, drone-derived images are used for training the proposed frameworks. For this purpose, an experimental plot was setup in the summer of 2015 in the Evogene Ltd. Farm (31.8833 N, 34.8437E, 80 m above mean sea level) near the city of Rehovot, in the coastal plain of Israel. Twenty commercial maize hybrids were planted in a total of 160 plots. The images were obtained on July 27, 2015. The center of the field was planted with sweet corn, and all plots were surrounded by additional corn plants to avoid border effects. The UAV was mounted with a 12-camera structure (MiniMCA12 of Tetracam Inc. Chatsworth, CA, USA) which captured the spectral information on a detector with 1280×1024 pixels. Eleven of the cameras captured the spectral bands centered at 420, 440, 490, 550, 640, 670, 700, 740, 780, 860, and 910 nm [22]. The twelfth camera, an incident light sensor (ILS), served as a reference for estimating the relative reflectance values. The inter-band geometric registrations of these UAV images were conducted using ground truth points and the feature shapes. In addition to these datasets, standard hyperspectral airborne datasets are also employed for comparing the proposed approaches with the existing ones.

3.2. Training and testing dataset generation

The UAV images as well as standard airborne datasets are classified using SVM to generate high-resolution classified maps. Also, these images are down-sampled (using bilinear, bicubic, and nearest neighbor) at different scales to yield the corresponding coarser resolution versions. The coarser images and their corresponding high-resolution classified maps are used for training and testing the sub-pixel classification frameworks. Similarly, the coarser images and their corresponding high-resolution images are used for training and testing the super-resolution frameworks. A subset of the available datasets is augmented to simulate various distortions and blurs prevalent in the drone-derived images. The main issues with the drone-derived images, considered in this study, are the distortions due to the pitch, roll and yaw vibrations. Furthermore, changes in atmospheric conditions or flight-height cause alternations in the scale as well as the spectral, spatial and radiometric properties. In this regard, data transformations such as translation, rotation, scaling, Gaussian noise, different blurring and smoothing operations, minor inter-band misalignments, and generative adversarial network based augmentations [2,3,18] are employed. These augmentations further help to increase the number of training and testing samples, and also improve the generalization capability of the network. A few of the augmentation and downscaling strategies adopted in this study are illustrated in Fig. 1. It may be noted that, similar to other supervised learning strategies, the network needs to be retrained when using for an entirely different topography as the approach proposed here is an offline one. However, transfer learning based approaches can be employed to facilitate such adaptations.

3.3. Sub-pixel classification

The proposed sub-pixel classification approach is discussed in detail in the following sub-sections. The Section 3.3.1 presents the overall sub-pixel mapping algorithm while Sections 3.3.2 and 3.3.3 discuss the proposed architectures and the loss functions respectively.

3.3.1. Algorithm

The soft-SVM-based unmixing, proposed by Arun and Buddhiraju [4], is adopted to transform the input low-resolution image to fractional abundance images (F_k for each class k). It may be noted that, in order to reduce the inter-band geometric registration errors, the soft-classification is implemented using higher level features (obtained from a trained CNN network). The rank images (R_{ck}) are the finer resolution approximations of the input low-resolution images. For each class k, the unconstrained rank image R_k is obtained from the corresponding fractional abundance image (F_k) by computing the weights (W_{ik}) at each sub-pixel location i as:

$$W_{ik} = \sum_{j=1}^{M} F_k(j) e^{-(D_{ij}+G_{ij})} \quad (1)$$

where M is the number of pixels in the neighborhood, $F_K(j)$ is the fractional abundance of kth class at jth coarse pixel, G_{ij} is the feature-space distance between the value of the jth coarse pixel and the coarse pixel corresponding to the ith sub-pixel, and D_{ij} is the distance between the centroid of the jth coarse pixel and the sub-pixel position (i). The initial unconstrained rank image (R_K), thus obtained, is then transformed to constrained rank image (R_{Ck}) by setting the top ranked Q_{jk} sub-pixels of each coarse

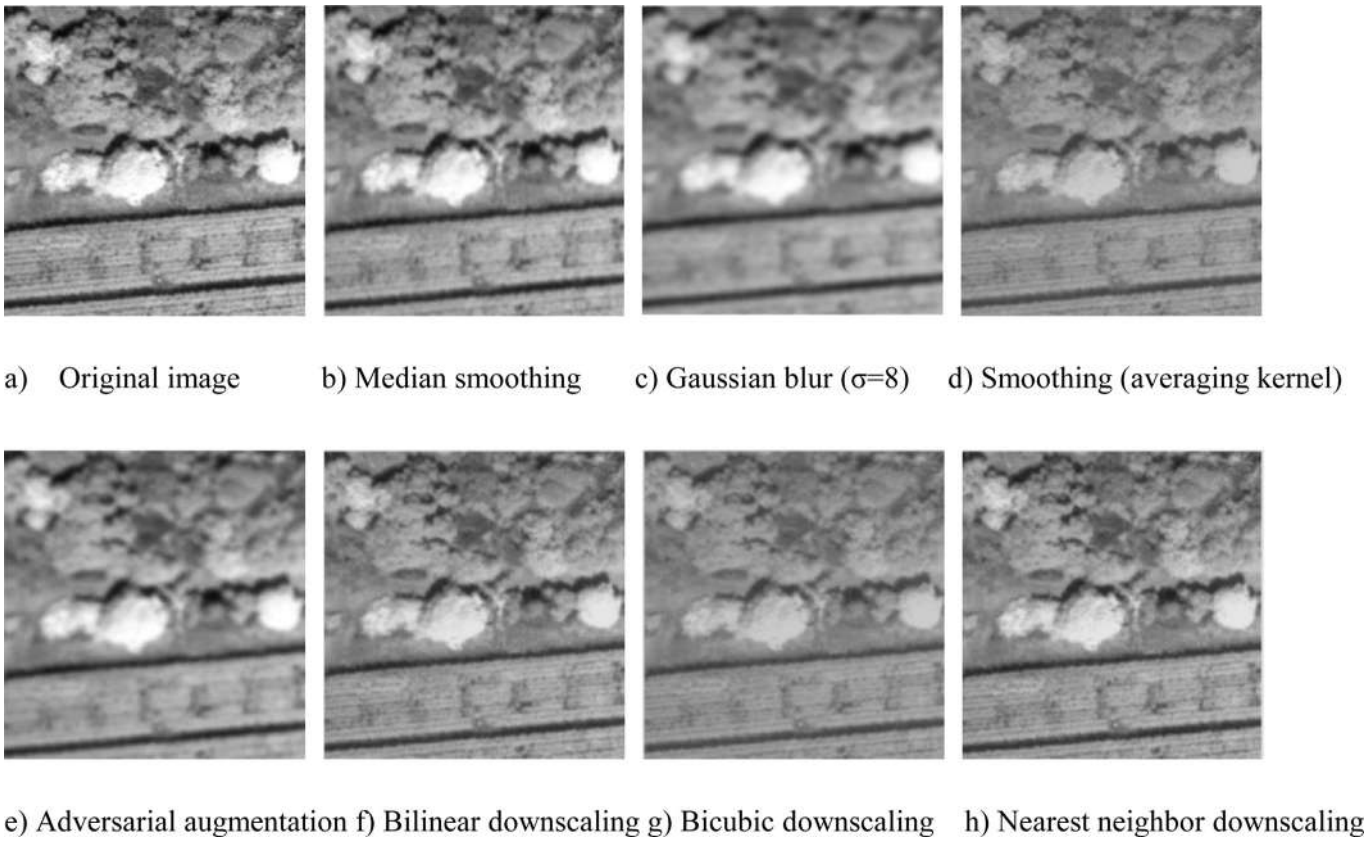

Fig. 1. Illustration of a few blurring, augmentation, and downscaling (zoom factor = 3) operations adopted to train the proposed framework.

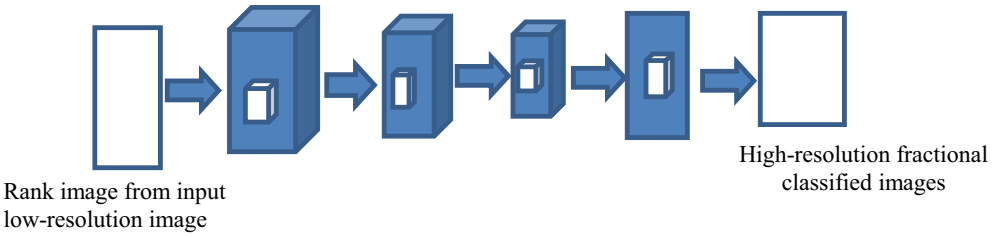

Fig. 2. Convolutional neural network architecture-1for sub-pixel mapping.

pixel (j) to 1 and the rest to zero. Here, Q_{jk} is estimated as: $Q_{jk} = (Z)^2 \times F_K(j)$ where Z is the zoom factor. The rank image (R_{ck}) derived from the simulated coarse image (Y) and the corresponding high-resolution fractional class image (A_K) derived from the original high-resolution image (W) are used to train the network. In other words, for each class k, the training pairs are of the form: (R_{ck}, A_k).

Once the network is trained, the rank image (R_{ck}') derived from input coarse image (to be tested), is fed to the trained CNN to yield the corresponding high-resolution fractional class image (I_K). Further, the distribution of labels in I_K is refined using the compatibility ($C(\mu_i,\mu_j)$) between each pair of classes μ_i and μ_j, which is computed as:

$$C(\mu_i, \mu_j) = P(\mu_i|\mu_j) \forall \mu_i, \mu_j \in C_L \quad (2)$$

where $P(\mu_i|\mu_j)$ is the conditional probability of the classes μ_i and μ_j in the coarse image, and C_L is the set of all classes. The labels whose average compatibility over the neighborhood is below a threshold are replaced by the most frequently occurring ones in their proximity.

3.3.2. Architecture

As discussed, the architectural choice for sub-pixel classification should consider the feature interpolation as well as the generalization of spatial distributions across scales. To start with, a simple feed-forward four-layer CNN (Fig. 2), that projects the input rank image to corresponding finer-scale fractional classified maps, is adopted. The random-search-based hyper-parameter optimization approach, proposed by [8], is adopted to find the appropriate values of various network parameters. In this regard, the optimal kernel sizes are found to be 5×5 in the first, 3×3 in the second, 1×1 in the third, and 5×5 in the fourth layer. Similarly, the optimal number of filters in the first, second, third, and fourth layers are set to be 32, 24, 12, and 1, respectively. Experiments, on various datasets over different scale factors, indicate that the architecture does not give satisfactory results.

The artifacts resulting from the lack of regularization, in the above architecture, is resolved by introducing an additional stream for learning the features. The refined architecture is shown in Fig. 3. The additonally introduced feature-stream (stream-1), reduces both the convergence time and artifacts, and also provides the required regularization. Here, stream-1 comprises a convolution-deconvolution network for super-resolving the fea-

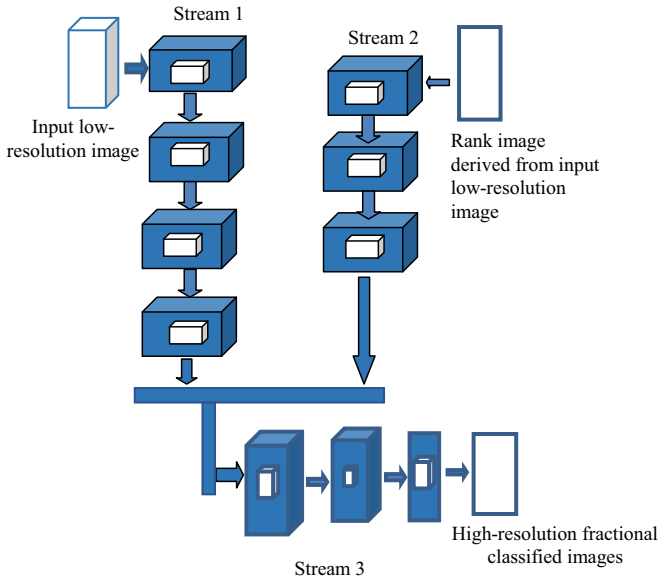

Fig. 3. Convolutional neural network architecture-2 for sub-pixel mapping.

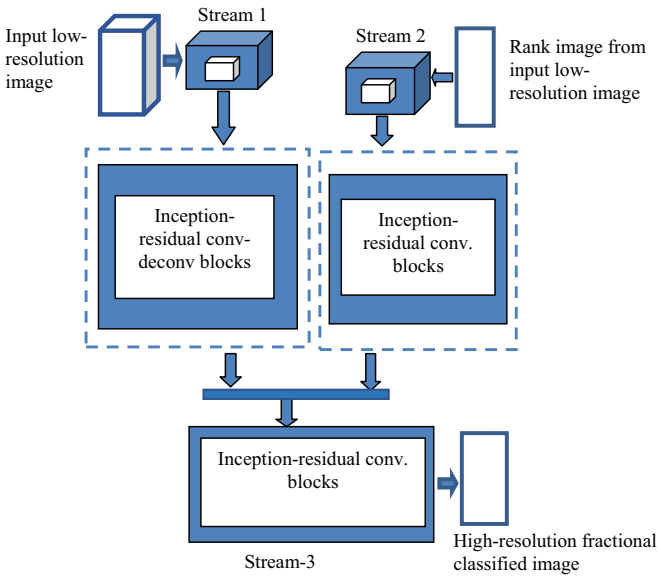

Fig. 4. Convolutional neural network architecture-3 for sub-pixel mapping.

tures (such as convolutional features or histogram of oriented gradients or edges), in parallel with the regular stream (stream-2), so as to ensure proper interpolation. It may be noted that, in this work, the convolutional features derived from the input coarse image are used as regularizing features. Among the four layers in stream-1, the first two are convolution layers, and the rest are deconvolution layers. In stream-2, the rank image derived from fractional coarse image is subjected to a series of convolutions. Finally, outputs from both the streams are stacked and inverted using stream-3 convolution filters to generate the required finer-scale classfied maps (I_k). The network uses the mean squared error between I_k and A_k as the loss function, which is minimized using the standard backpropagation. Also, leaky rectified linear units (ReLu) [68] are used as activation functions in which the leaky parameter is tuned and set to 1/3.

It may be noted that instead of deriving features from the upscaled coarse image, here the upsampling of input features is implemented indirectly using the convolution-deconvolution stream (stream-1). This approach improves the accuracy and considerably reduces the artifacts. The deconvolution is formulated as upsampling followed by convolution, and upsampling is implemented by replacing each value with a block having original value in the center and zero in all the other entries. The convolution-deconvolution stream (stream-1) has kernel sizes of 3×3 in the first, 3×3 in the second, 5×5 in the third, and 5×5 in the fourth layer. The numbers of filters in these layers are set to be 12, 24, 12, and 6, respectively. Similarly, the convolution stream (stream-2) has kernel sizes of 5×5 in the first, and 3×3 in both second and third layers. The filter numbers are 128, 64, and 32, respectively. The stream-3 convolution network has kernel sizes of 3×3, 1×1, and 5×5 in the first, second and third layers, respectively. The 1×1 layer provides non-linear mapping from low-resolution representation to high-resolution ones. The numbers of filters in these layers are set to be 32, 24, and 1, respectively. As in the case of architecture-1, here also the optimal network parameters are selected using the hyper-parameter optimization proposed by Bergstra et al. [8].

Although the learning capability is improved, the required optimal network-depth is found to be sensitive to both the scale factor, and the scene complexity. Furthermore, the convergence time is very high for high scale factors ($Z > 2$). In this regard, the

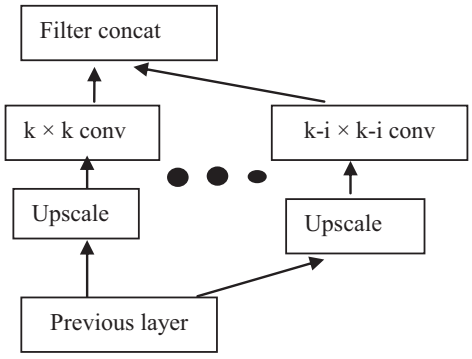

Fig. 5. Basic deconvolution block for stream-1 of architecture-3 (given in Fig. 4).

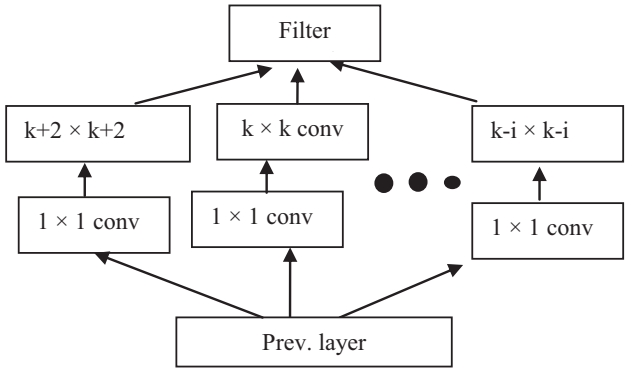

Fig. 6. Basic convolution block for streams-1 and 2 of architecture-3 (given in Fig. 4).

simple convolution and deconvolution units, used in the previous architectures (architectures-1 and 2), are replaced by inception-residual blocks. The resulting architecture simulates the inception networks and is shown in Fig. 4. Instead of directly adopting the inception-residual block [54], which is meant for classification, it is modeled for streams-1, 2, and 3 as presented in Figs. 5–7. These blocks constitute multiple convolutional units with different-sized kernels. The diversity of kernels increases the learning capability and also reduces the required optimal depth (as compared to architectures-1 and 2). Both the kernel size (k & i) and the num-

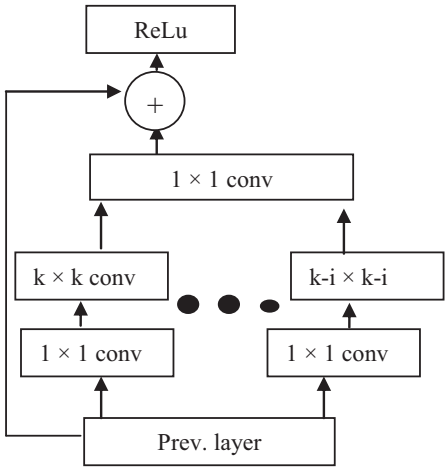

Fig. 7. Basic convolution block for stream-3 of architecture-3 (given in Fig. 4).

ber of filters in each unit are changed accordingly for each layer. In the whole implementation, i varies from k to 3. As in the case of architecture-2, a four-layer convolution-deconvolution network (stream-1), with k as 7, 7, 5, and 5, respectively, is used. The number of filters in the convolution and deconvolution units of first, second, third and fourth layers are set to 6, 8, 6, and 6, respectively. The stream-2 is formulated as a three-layer convolution stream with $k=5$ in all layers, and the numbers of filters are set to 6, 8, and 6, respectively. For stream-3, a three-layer network is used with the number of filters in the respective blocks set as 8, 8, and 1. The settings are mostly similar to architecture-2 for ease of comparison.

Generally, for sub-pixel classification, cross-entropy based loss functions are minimized to finetune the network. In this study, an additional loss function (L_{MC}) is proposed to address the misclassification, i.e.,

$$L_{MC} = 10^{-1} \times no : of\,(values > 1) in \left(\sum_{C=1}^{E} \mathbf{F_C} \right) \tag{3}$$

where F_c is the reconstructed fractional classified image of cth class, and E is the number of classes. The proposed misclassification loss helps to prevent the same sub-pixel from being assigned to different classes.

3.4. Super-resolution

This section discusses the proposed super-resolution algorithm (Section 3.4.1) and also details the evolution of the proposed architecture (Section 3.4.2). It may be noted that the algorithm is generic but the architecture is specific to the multispectral/hyperspectral drone-derived/airborne datasets.

3.4.1. Algorithm

The simulated coarse image (Y) and its corresponding fine resolution version (W) are used for training the super-resolution framework, i.e., here the training pairs are of the form (Y, W). The trained network is then fed with the input coarse image (Y') so that each band is reconstructed to generate the high-resolution image (W'). It may be noted that in the proposed approach, all the bands are reconstructed simultaneously considering the spectral aspect of the image. This simultaneous learning yields better results than the individual-band-based enhancement. Also, unlike the previous approaches, upsampling is performed within the convolution-deconvolution network rather than using a bilinear version of the input.

3.4.2. Architecture

The network architectures for super-resolution are similar to those for the sub-pixel classification. The simple layer architecture similar to that shown in Fig. 2 did not yield the expected results (results were smoother and blurrier due to lack of regularization). Hence, the three-stream architecture similar to that in Fig. 3 is used. However, in contrast to the sub-pixel classification, two convolution-deconvolution streams (stream-1&2) are used in parallel followed by a convolution stream (stream-3). In the current implementation, the network parameter settings are tuned for 11 bands using the hyper-parameter optimization discussed in Bergstra et al. [8]. The streams-1 and 2 convolution-deconvolution units have kernel sizes of 7×7 in the deconvolution phase and 5×5 in the convolution phase. The number of filters in these layers are set to be 32, 64, 128, 64, 32, and 16, respectively. Similarly, the stream-3 convolution layers have kernel sizes of 5×5 pixels in the first, 1×1 pixels in the second, and 5×5 pixels in the third layer. The number of filters in these layers are set to be 11.

The stream-1 convolution-deconvolution units transform the features to a finer spatial-scale. Similarly, the stream-2 convolution-deconvolution units are employed to project the coarser image to a target-scale. Further, the stream-3 convolution units invert the upsampled feature (from stream-1) and perceptual image representation (from stream-2) to the required super-resolved image. The inversion implemented using streams-1, 2, and 3 facilitate the mapping of features, as well as the color information, to a finer spatial-scale. Also, since the proposed inversion is implemented in a convolutional latent space, the approach can better model the underlying manifold of a given set of data. Unlike individual band enhancements, here all the filters convolve all the bands, thus enabling multidimensional learning, improving the spectral fidelity. Moreover, in order to minimize the effect of inter-band geometric mismatches (due to the unstable UAV platform), the inter-band mixing is proposed at higher layers.

Although the approach yields satisfactory results, it introduces artifacts for higher scale factors. Moreover, learning is found to be dependent on the network depth, thereby affecting the computational efficiency. These issues can be resolved by using the inception-residual blocks, given in Figs. 5–7, instead of simple convolution and deconvolution units. In this regard, the three-stream architecture (similar to the one in Fig. 3) is refined to yield the final super-resolution architecture (architecture-3) presented in Fig. 8. The basic inception-residual blocks are similar to those shown in Figs. 5–7, except that both stream-1 and stream-2 are convolution-deconvolution streams. In this study, four-layer stream-1 and stream-2 networks are used with the value of k set to five in each. Also, the numbers of filters are set to be 64, 128, 64, and 32. For the stream-3 of depth three, k is set to five and the numbers of filters are set to 16, 8, and 11. The approach considerably reduces the artifacts and improves the perceptual appearance at a lower number of iterations. An ensemble version of the proposed framework, inspired by Wang et al. [61], is simulated by improving the kernel diversity and randomly selecting a subset of kernels.

4. Analysis of the super-resolution framework

In this section, various proposed super-resolution architectures, as well as the state-of-the-art techniques, are compared. In addition, variation in the accuracy of the proposed method with respect to the network parameters is analyzed. The average RMSE computed between each reconstructed band and its original high-resolution version is used as a performance measure. However, as discussed previously, the perceptual resemblance cannot be expressed using these measures. Hence, the kappa statistics

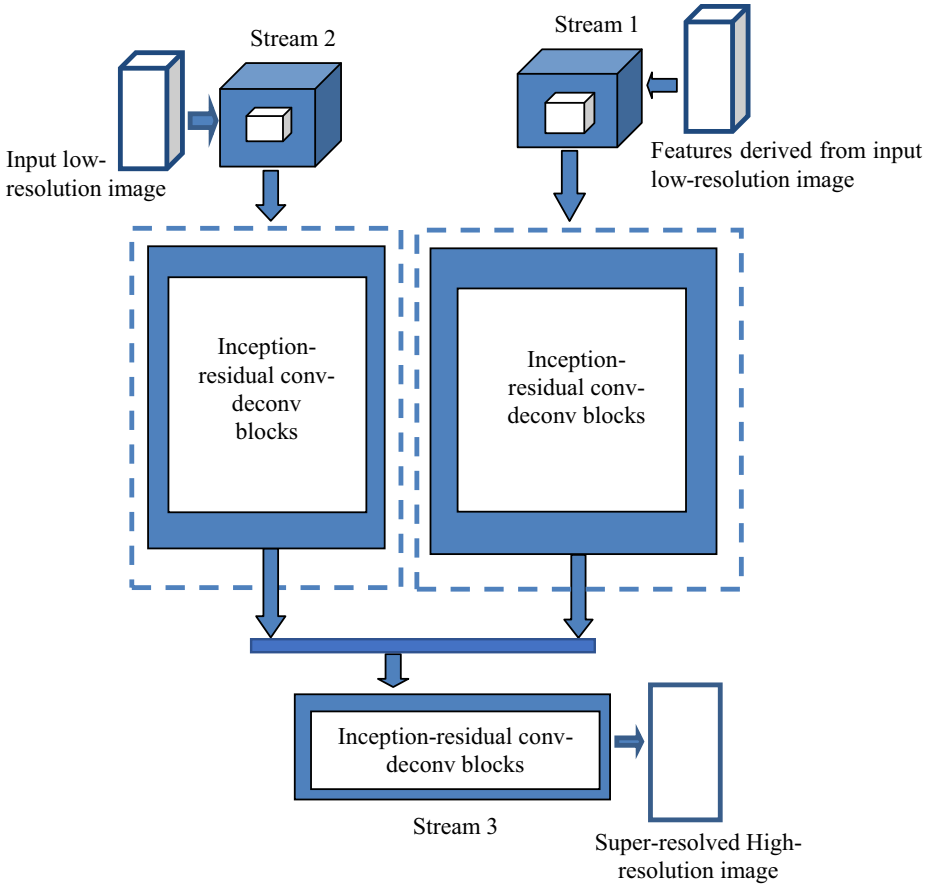

Fig. 8. Convolutional neural network architecture-3 for super-resolution.

Table 1
Comparison of super-resolution architectures.

Method	Scale	RMSE	Kappa statistics	MS-SSIM	FSIM	Overall accuracy (%)	Exec. time (sec.)
Arch-1	2	0.31	0.96	0.8419	0.8136	96.26	42
	3	0.37	0.96	0.8031	0.7617	95.67	78
	4	0.54	0.94	0.7841	0.7205	93.85	113
Arch-2	2	0.19	0.97	0.8718	0.9161	98.53	94
	3	0.25	0.97	0.8404	0.8963	97.82	119
	4	0.32	0.96	0.8239	0.8870	96.71	156
Arch-3	2	0.14	0.99	0.9176	0.9218	99.45	123
	3	0.20	0.98	0.9093	0.9056	98.28	197
	4	0.39	0.98	0.8865	0.8934	97.65	251

and overall accuracy of the classified maps (generated from the reconstructed images), along with the multiscale structural similarity index measure (MS-SSIM), and the feature similarity index measure (FSIM), are employed. Throughout the experiments, the mini-batch size for training is set to 200; momentum and weight decay for the backpropagation are set to 0.8 and 10^{-3} respectively (obtained through cross-validation); the learning rate is initially set to 0.6 and is depreciated by a factor of 4 after every 50 epochs. All the models are analyzed for 200 epochs.

4.1. Architecture

Comparison of the three proposed super-resolution architectures is summarized in Table 1. The network parameter settings corresponding to these results are in accordance with the Section 3.4.2. Although architecture-2 gives good results, both the convergence time, and the perceptual appearance of the results of architecture-3, are better. Repeated experiments over different scale factors, and different network parameter settings, yielded similar trends. The ensemble-based modification of the proposed approach improves the results but at the cost of increased running time. It is found that the cascaded implementation of the super-resolution models, for higher zoom factors, is not suitable for multispectral UAV images; hence the proposed approach of deconvolution should be preferred.

4.2. Loss functions

A summary of the analysis of the loss functions is presented in Table 2. It may be noted that the presented results are for architecture-3, while other architectures also yield similar results. The generative-adversarial and perceptual domain based loss functions gave better results than the conventional MSE-based approaches. However, the computational expenses of the former are higher when compared to the latter. The effects of these functions are better reflected in the values of MS-SSIM, and FSIM

Table 2
Analysis of loss functions for proposed super-resolution method.

Method	Scale	RMSE	Kappa statistics	Overall accuracy (%)	MS-SSIM	F-SIM	Exec. time (sec.)
RMSE	2	0.16	0.98	98.14	0.9146	0.9207	103
	3	0.24	0.96	97.49	0.9028	0.9041	147
	4	0.41	0.95	96.21	0.8816	0.8897	211
Peak signal to noise ratio (PSNR)	2	0.12	0.98	99.49	0.9217	0.9296	147
	3	0.21	0.98	98.37	0.9104	0.9104	212
	4	0.35	0.97	97.72	0.8919	0.9012	234
Perceptual loss using visual geometry group network features	2	0.08	0.99	99.73	0.9348	0.9459	191
	3	0.17	0.98	98.52	0.9286	0.9281	234
	4	0.24	0.97	97.96	0.9050	0.9204	280
Generative-adversarial loss	2	0.06	0.99	99.82	0.9468	0.9497	221
	3	0.15	0.98	99.06	0.9315	0.9346	266
	4	0.19	0.98	98.57	0.9208	0.9217	307

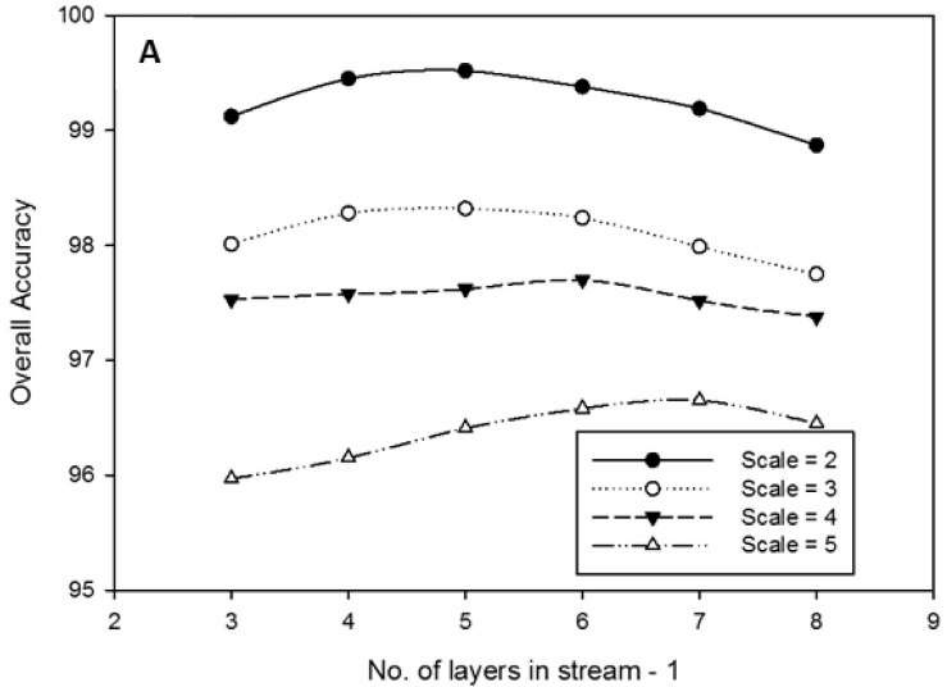

Fig. 9. Illustration of variation in accuracy with number of layers in stream-1 for super-resolution architecture-3 (given in Fig. 8).

than RMSE. Hence, it can be concluded that the overall accuracy, kappa statistics, MS-SSIM, and FSIM are better estimates for perceptual accuracy. Also, experiments over different datasets indicate that the use of spectral dissimilarity loss (inverse of the average spectral similarity between the original and the reconstructed images) along with the above loss functions significantly improve the results.

4.3. Depth of the networks

The analysis of various architectures over network depth reveals that both the performance, and the accuracy, vary with the depth. Specifically, for architecture-1, the convergence time and the accuracy are highly sensitive to the number of layers. However, architecture-2 and 3 give comparatively better stability. In all the architectures, an increase in the number of convolution-deconvolution layers improves the accuracy to a limit beyond which it gradually deteriorates. The depreciation is much slower in architecture-3 when compared to the others. This can be attributed both to the use of multi-sized kernels (kernel diversity) and to residual learning, which improves the learning capability. The increase in the depth of stream-3 only slightly improves the accuracy which soon reaches saturation. Generally, the computational expenses increase exponentially with the increase in network depth. The variation in accuracy with the depth of different network-streams for architecture-3 is illustrated in Figs. 9 and 10. It is observed that the optimal depth depends on the scale factor, scene homogeneity, as well as the number of bands for which the network is tuned.

4.4. Size and number of filters, diversity of kernel sizes

In all the architectures, an increase in the number of filters improves the accuracy but increases the running time. Hence an optimal choice depends on the tradeoff between both. However, in contrast to the earlier findings reported by Dong et al. [12] and Liebel et al. [39], the increase in the filter size (particularly in streams-1 and 2) improves the accuracy only to a certain limit. This can be attributed to the inability of the network in learning smaller features. Hence, the optimal filter size should be determined with reference to the scene homogeneity/complexity. Analysis of the variation in accuracy of the proposed framework (architecture-3), with respect to the size and the number of filters, is summarized in Figs. 11 and 12. Experiments also indicate

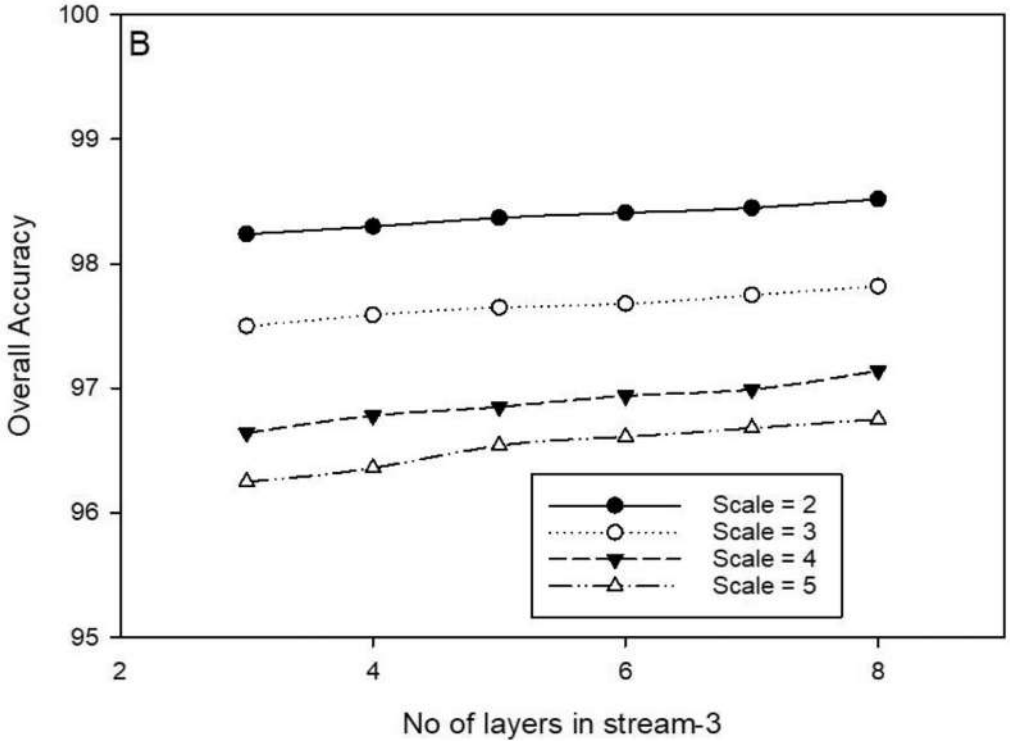

Fig. 10. Illustration of variation in accuracy with number of layers in stream-3 for super-resolution architecture-3 (given in Fig. 8).

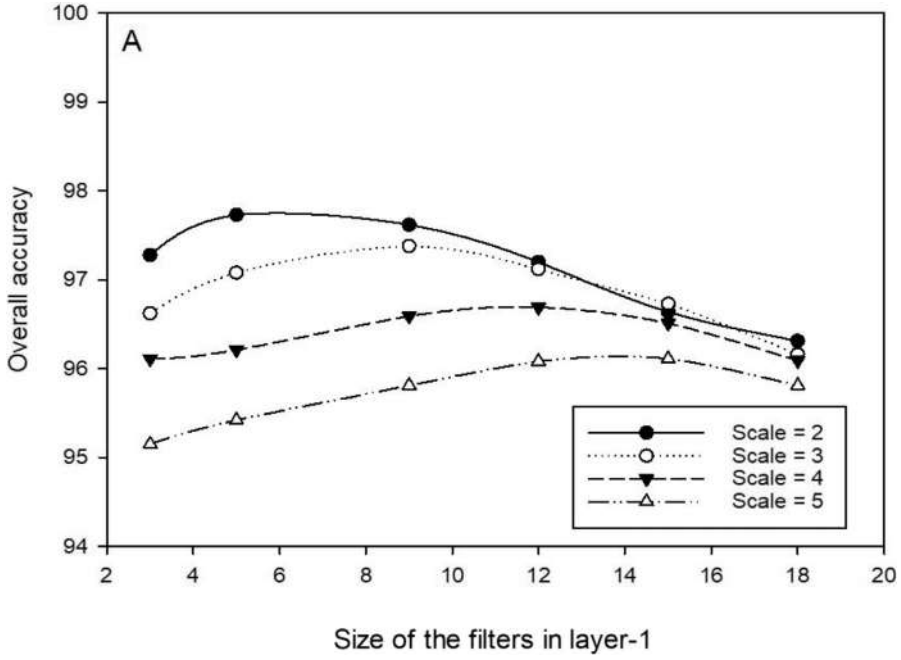

Fig. 11. Illustration of variation in accuracy with size of filters for super-resolution architecture-3 (given in Fig. 8).

that the increase in the diversity of kernel sizes improves the accuracy and yields better stability and reconstruction, even at shallower depths.

4.5. Regularizing features

All the architectures show a similar trend with the increase in the number of regularizing features, i.e., both the accuracy and the execution time increase almost exponentially. The trend of the variation in accuracy for architecture-3 is depicted in Fig. 13. It may be noted that the number of filters and the kernel diversity may be increased to maintain the pace of these improvements. From Table 3, it is evident that, among the various features, histogram of oriented gradients provides the best reconstruction, followed by Gabor.

4.6. Fine-tuning of existing networks

The available networks (super-resolution convolutional neural networks (SRCNN), visual geometry group network, and

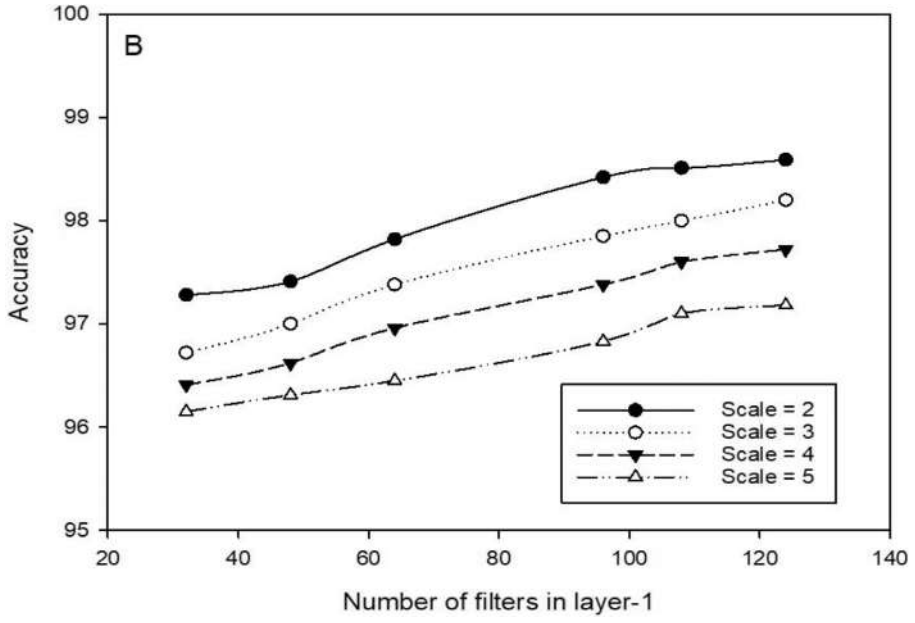

Fig. 12. Illustration of variation in accuracy with size of filters for super-resolution architecture-3 (given in Fig. 8).

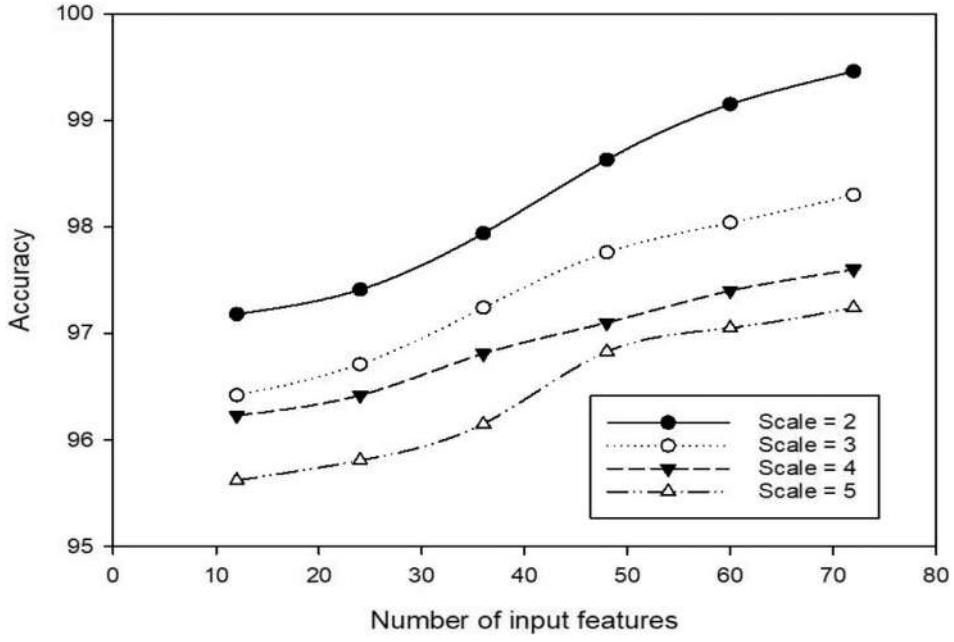

Fig. 13. Illustration of variation in accuracy with number of input features for super-resolution architecture-3 (given in Fig. 8).

Table 3
Analysis of input features for super-resolution.

Method	Scale	RMSE	Kappa statistics	Overall accuracy (%)	MS-SSIM	F-SIM	Exec. time (sec.)
Edges	2	0.14	0.99	99.45	0.9176	0.9218	123
	3	0.20	0.98	98.28	0.9093	0.9056	197
	4	0.39	0.98	97.65	0.8865	0.8934	251
Histogram of oriented gradients	2	0.05	0.99	99.53	0.9404	0.9347	184
	3	0.09	0.98	98.84	0.9239	0.9183	217
	4	0.13	0.98	98.17	0.9063	0.9027	253
Gabor Filters	2	0.07	0.98	99.21	0.9298	0.9301	162
	3	0.12	0.97	98.05	0.9140	0.9124	199
	4	0.18	0.96	97.64	0.8963	0.8869	234
Harris Corner Detectors	2	0.12	0.98	99.18	0.9057	0.9017	135
	3	0.15	0.96	97.93	0.8941	0.8943	197
	4	0.26	0.96	97.21	0.8706	0.8862	242

Table 4
Comparative analysis of the proposed super-resolution method.

Method	Scale	RMSE	Kappa statistics	Overall accuracy (%)	MS-SSIM	F-SIM	Exec. time (sec.)
Proposed	2	0.18	0.99	99.45	0.9176	0.9218	123
	3	0.20	0.98	98.28	0.9093	0.9056	197
	4	0.39	0.97	97.44	0.8865	0.8934	251
Proposed (ensemble version)	2	0.12	0.99	99.51	0.9340	0.9451	218
	3	0.14	0.99	99.04	0.9168	0.9236	274
	4	0.26	0.98	97.89	0.8901	0.9068	316
SRCNN* (Liebel et al., [39]	2	0.85	0.94	96.29	0.8826	0.9022	269
	3	1.04	0.91	93.48	0.8591	0.8968	332
	4	1.57	0.87	92.56	0.8315	0.8324	417
Wang et al. [61]	2	0.21	0.98	98.17	0.9117	0.9106	315
	3	0.28	0.98	97.93	0.8964	0.8851	382
	4	0.32	0.97	97.76	0.8738	0.8680	429
Dong et al. [13]	2	0.65	0.93	94.45	0.8429	0.8739	96
	3	1.28	0.92	92.29	0.8134	0.8516	152
	4	1.73	0.90	89.61	0.7910	0.8329	205

GoogLeNet) cannot be modeled directly for the proposed super-resolution framework. The remodeled SRCNN, with only the stream-3 architecture, does not yield the expected results. In addition, fully convolutional versions of the conventional networks are also investigated. Although these networks perform better than the modified SRCNN, they still are not comparable with the proposed one. The individual reconstruction of bands using these networks is also tested; however, the results are worse than those of the previous experiments. Retraining of the proposed networks with the MIT place dataset [78] shows a better performance at lower scale factors.

4.7. Comparison with the state-of-the-art

A comparative analysis of the proposed approach (architecture-3) with the state-of-the-art is summarized in Table 4. An illustration of these results is presented in Fig. 14. As is evident, the proposed method gives better accuracy. It is observed that the simultaneous learning of bands better enhances the reconstruction capability, in contrast to the individual band enhancements. Also, when compared to the conventional approach of using an upsampled image as input to the CNN, the proposed approach of using a higher-level feature space, for upsampling, improves the results. Most importantly, the sparse-coding-based strategies give limited results for datasets in which the classes (objects) lack regular spatial structures. The proposed framework projects the images to finer spatial grids even when the sparse-coding assumptions are not satisfied. Furthermore, the approach removes the blur as well as other distortions prevalent in the UAV images. The multi-sized kernels, used in the proposed architectures, effectively model the diverse features and also improve the sparse representations. This further helps to recover the mixed features which even the sparse-representation-based approaches are not able to model. The higher Kappa and F-SIM values of the proposed method, presented in Table 4, indicate that the mixed pixels are properly super-resolved.

The UAV datasets used in this experiment cover the vegetation and urban classes, and the proposed approach is found to be effective for such land covers. The better performance of the approach can mainly be attributed to the regularized inversion, which models the mapping of spatial features as well as the color information. Although an ensemble-based modification of the proposed framework, based on Wang et al. [61], improved the results, its applicability is limited for inputs with fewer spectral bands. However, parallelization of different streams can be exploited to considerably reduce the running time.

The absence of standard UAV datasets limits the comparative analysis of the proposed approach with prominent approaches. However, to illustrate the better performance of the proposed approach, a comparative evaluation of the same over standard datasets is presented in Table 5. The average RMSE, Kappa, MS-SSIM, and F-SIM values on airborne datasets such as Indian Pines, Salinas, Cuprite, Pavia and KSC illustrate that the approach performs better over various types of data. The sparse-coding-based approaches often suffer from bottleneck as the spectral dimension of the input increases; however, the proposed framework scales well, even for hyperspectral datasets. Experiments indicate that the proposed framework better preserves the spectral fidelity (better value of spectral similarity measure in Table 5) of the super-resolved results when compared to the conventional approaches. This illustrates the capability of the proposed spectral information based loss function as well as the feature-guided inversion in properly mapping the color information. Currently, like any other supervised approaches, the proposed framework needs to be retrained in order to be employed for an entirely different terrain. However, advanced data augmentation and multiple downscaling strategies, discussed in Section 3.2, improves generalization capability. Drone-derived images from different terrains, covering more land cover classes other than vegetation, can be used to analyze the universality of the approach. The spectral-spatial prior derived from the coarse image can be employed to refine the super-resolution result to improve the universality. In addition, neighborhood-frame-based transfer learning strategy can be explored as well.

5. Analysis of the sub-pixel classification framework

In this section, the performance of the proposed sub-pixel mapping framework (discussed in Section 3.3) is analyzed in the context of drone-derived images. Throughout the experiments, the mini-batch size for training is set to 200; momentum and weight decay for the backpropagations are set to 0.4 and 10^{-3} respectively (obtained through cross-validation); the learning rate is initially set to 0.7 and is depreciated by a factor of 4 after every 50 epochs. All the models are analyzed for 200 epochs.

5.1. Comparison with the state-of-the-art

The proposed approach (architecture-3) is compared with the state-of-the-art sub-pixel classification methods over different scale factors (3, 2, and 4), and the results are summarized in Table 6. The high-resolution fractional classified map of each class is compared with its corresponding ground truth to estimate the RMSE. Also, to better analyze the perceptual resemblance, the average classification accuracies (in terms of Kappa statistics and overall accuracy) are also estimated. As is evident from Table 6, the

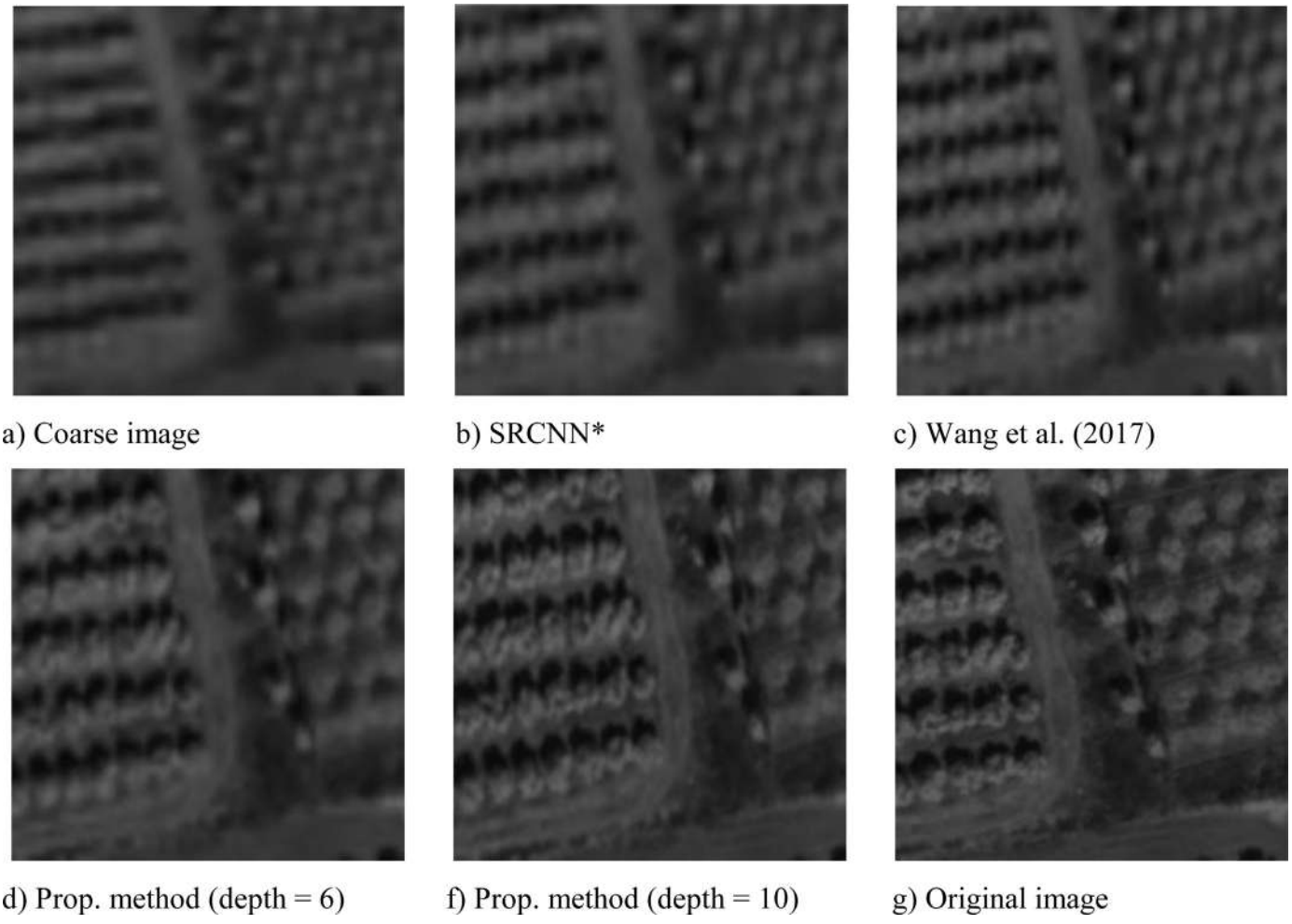

Fig. 14. Results of proposed super-resolution architecture-3 (given in Fig. 8).

Table 5
Comparative analysis of the proposed super-resolution method over standard airborne datasets.

Method	Scale	PSNR	MS-SSIM	Kappa statistics	Avg. spectral similarity	Exec. time (sec.)
Bi-cubic	2	19.84	0.7356	0.79	0.8742	28
	3	17.62	0.7118	0.77	0.8685	32
	4	16.21	0.7265	0.73	0.8503	41
Dong et al. [13]	2	31.91	0.8651	0.92	0.9480	280
	3	28.35	0.8439	0.90	0.9472	318
	4	26.78	0.8394	0.89	0.9464	374
Fu et al. (2017)	2	33.90	0.9205	0.94	0.9591	239
	3	32.66	0.9076	0.94	0.9537	278
	4	31.07	0.8897	0.92	0.9518	306
Proposed approach	2	37.12	0.9847	0.97	0.9905	123
	3	35.81	0.9315	0.96	0.9872	197
	4	35.04	0.9214	0.96	0.9816	251

Table 6
Comparative analysis of the proposed sub-pixel classification method.

Method	Scale	RMSE	Kappa statistics	Overall accuracy (%)	Exec. time (sec.)
Proposed method	2	0.07	0.98	99.16	58
	3	0.13	0.97	97.12	91
	4	0.17	0.97	97.05	123
Indicator Co-Kriging based method [6]	2	0.26	0.92	92.65	39
	3	0.34	0.89	91.81	65
	4	0.42	0.86	88.13	104
CNN based method [7]	2	0.10	0.97	98.05	79
	3	0.19	0.96	96.44	137
	4	0.28	0.95	93.96	186
CNN method [5]	2	0.12	0.96	97.31	71
	3	0.24	0.94	96.38	118
	4	0.35	0.91	92.16	162

Table 7
Class wise accuracy comparison.

#	Class	No. of samples	Avg. accuracy
1	Road	6631	98.12
2	Meadows	18,649	99.07
3	Agriculture land	2099	98.91
4	Plantations	3064	99.43
5	Forest	1345	96.06
6	Bare Soil	5029	98.71
7	Water	1330	100.00
8	Vehicle	100	99.94

proposed framework outperforms the prominent sub-pixel mapping approaches. While the other approaches merely optimize the spatial contiguity, the proposed framework models the mapping between coarser and finer scale distributions of classes. In addition, it strives to interpolate the feature information. It should be noted that the architecture-3, adopted here, is not fully optimized towards computational performance. However, the execution time is still comparable with the other approaches, and parallelization can be explored to further reduce the same.

In order to further analyze the proposed approach, the average class-wise accuracies, over UAV datasets, are summarized in Table 7. For most of the classes considered, producer and consumer accuracies of the proposed approach are found to be better than the others. Significant improvements, in comparison with the existing approaches, are observed for the detection of urban as well as other topographical features. The better results can be attributed to the ability of the approach in capturing the spatial topologies and textures inherent in the land-cover classes. Also, the neighborhood consideration facilitates the resolution of blurred or distorted edges and other features. The boundaries of agricultural fields as well as the crop-separations are successfully captured from even blurred images. Also, the regularized inversion ensures the proper mapping of color information which is evident from the correct classification of even the classes having random patterns or textures (e.g. Forests and Meadows).

Due to the unavailability of standard UAV datasets, for comparative evaluation, the airborne datasets such as Indian pines, Salinas, Pavia, Cuprite and KSC were employed. The average kappa statistics, overall accuracy and Earth mover's distance for different prominent approaches, over these datasets, are summarized in Table 8. A graphical illustration of the comparison is provided in Fig. 15. As is evident, the proposed approach performs better when compared to the other approaches irrespective of the land cover. These experiments confirm the ability of the framework in simultaneously mapping the color and the feature information. In order to fine-tune the network for a slightly different terrain, compatibility-based refinements proposed in Arun et al. [6] can be adopted.

5.2. Network parameter setting

Network parameter settings, for sub-pixel classification architectures, show a trend similar to that of super-resolution. The increase in stream-3 depth improves the accuracy only slightly and leads to saturation. Although deeper stream-2 layers give better accuracy, a blind increase in stream-1 depth may adversely affect the reconstruction. For architecture-3, the gradient of depreciation in accuracy (with the increase in stream-1 depth) is very low, indicating better stability of learning. It may be noted that, for all the architectures, the optimal depth depends on the scale factor for which the network is tuned. Also, the computational expenses generally increase with an increase in the network depth.

Among the various architectures, architecture-3 gives better accuracy and convergence time. In all the architectures, an increase in the number of filters improves the accuracy but at the cost of running time. An increase in the size of stream-1 and 2 filters beyond a certain threshold, which is determined based on the scene complexity, adversely affects the accuracy. Although a similar increasing trend is observed for stream-3, it saturates after a threshold. The increase in number of regularizing features improves the accuracy, and more filters are needed to avoid the saturation. Among the various features, convolutional features give the best results followed by Histogram of Oriented Gradients, Gabor filters, Harris corner detectors, and edges, in descending order. It can be seen that the increase in filter diversity (number of different kernel sizes) improves the accuracy and enables better reconstruction at shorter depths. Furthermore, it also improves the convergence time. An analysis of the possible loss functions for sub-pixel classification is summarized in Table 9. As is evident, the proposed loss function improves the accuracy when compared to the conventional approaches. However, it causes a slight increase in the execution time; hence, an optimal choice depends on the tradeoff between computational performance and accuracy.

5.3. Fine-tuning of existing networks

Although conventional CNNs are mainly tuned for classification, recent SRCNN and fully convolutional networks [43] can be explored for sub-pixel mapping. Fully convolutional layers are directly adapted for stream-1, whereas tailored versions are used for the other streams. For stream-2 and stream-3, an extra layer is added to convolve the rank image. Retraining of this modified configuration does not yield promising results. A feasible alternative is to retrain the proposed network with the conventional datasets (Visual geometry group network, GoogLeNet, and others). This approach yields satisfactory results, especially for lower scale factors. The re-modeling of the SRCNN network [12] (setting $c = 1$), with only stream-3 layer, has also been investigated. Although the retraining of such a network (with the datasets used in this study) yields good results, the proposed approach is found to be far bet-

Table 8
Comparative analysis of the proposed sub-pixel mapping method on standard datasets.

Method	Scale	RMSE	Kappa statistics	Overall accuracy (%)	Exec. time (sec.)
RMSE	2	0.08	0.98	99.16	88
	3	0.13	0.97	97.12	131
	4	0.17	0.97	97.05	203
Cross entropy	2	0.07	0.98	99.29	106
	3	0.11	0.98	98.64	174
	4	0.16	0.97	97.68	210
Cross entropy + Misclassification loss	2	0.05	0.99	99.51	143
	3	0.10	0.98	99.29	186
	4	0.15	0.98	98.70	215

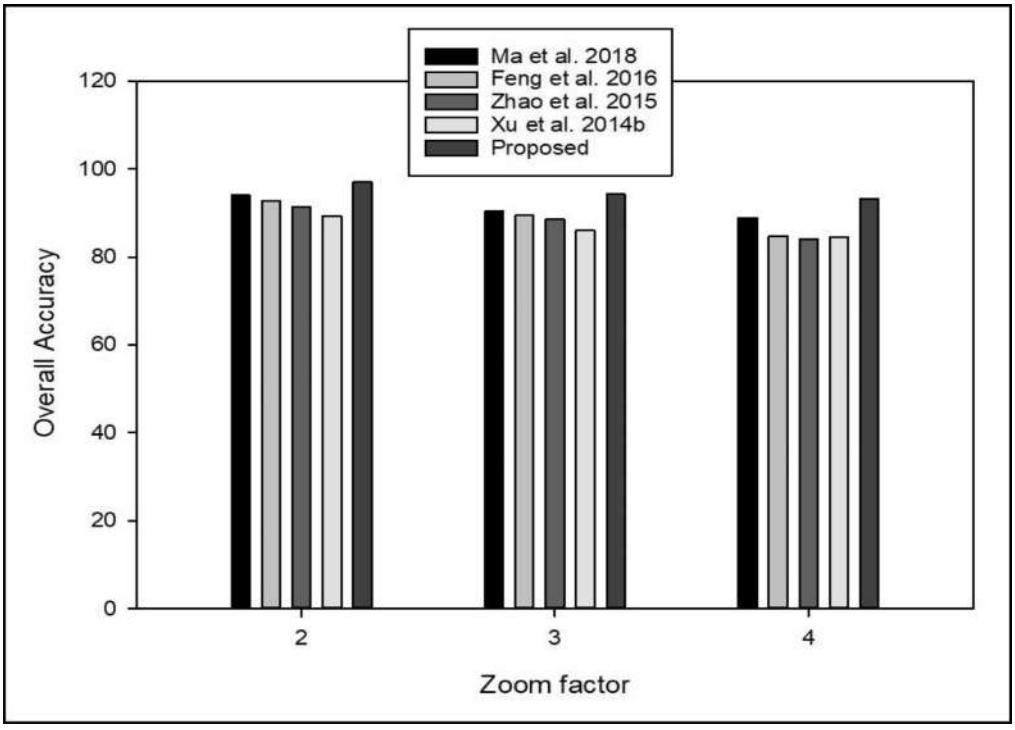

Fig. 15. Comparison of the overall accuracy of prominent sub-pixel mapping approaches on standard dataset.

Table 9
Comparative analysis of loss functions for sub-pixel classification.

Method	Scale	Avg. Kappa statistics*	Avg. Overall accuracy (%) *	Avg. Earth Mover's Distance	Exec. time (sec.) *
Ma et al. (2018)	2	0.91	94.02	2.89	132
	3	0.89	90.35	4.13	209
	4	0.85	88.91	4.97	287
Feng et al. [17]	2	0.90	92.76	3.74	51
	3	0.84	89.54	4.88	78
	4	0.81	84.80	5.20	124
Zhao et al. [76]	2	0.88	91.32	3.91	147
	3	0.86	88.65	4.07	181
	4	0.79	84.16	6.84	216
Xu et al. [70]	2	0.85	89.33	3.39	62
	3	0.82	86.07	5.12	108
	4	0.79	84.49	7.05	177
Proposed method	**2**	**0.94**	**97.09**	1.23	88
	3	**0.92**	**94.26**	**2.18**	131
	4	**0.90**	**93.28**	**3.94**	203

ter, especially at higher scale factors. This further illustrates the significance of convolution-deconvolution networks in learning the semantic aspects of the image.

6. Conclusion

This study analyses various architectural choices for super-resolution and sub-pixel mapping of drone-derived images, and in addition, proposes some improved frameworks that consider the specific characteristics of the data as well as the distortions prevalent in it. For inputs with increased spectral dimension and scene complexity, the sparse-code-based super-resolution approaches suffer from bottleneck and results in corrupted dictionaries and sparse codes. The feature-guided network-inversion, proposed in this study, addresses these issues and yield sharper and better reconstructions. It is observed that the approach gives 2–5% improvement in accuracy when compared to the prominent approaches such as Dong et al. [13], Wang et al. [61], and Feng et al. [17]. The significance of the approach is better observable with the increase in distortions or spectral dimension of the input datasets, where the improvement is above 10 percent. The increase in the number of regularization features, in the proposed approach, enhances the high frequency details in the super-resolved output, but at the cost of computation time. Also, this deblurring property resolves the effect of distortions prevalent in the UAV datasets. The initial bicubic/bilinear upscaling, adopted in earlier super-resolution approaches, causes reconstruction artifacts. Hence, in the proposed approach, the coarser input image is dynamically upscaled using convolution-deconvolution networks, resulting in an average accuracy improvement of 4–15% on noisy datasets, specifically, for higher scale factors. The proposed sub-pixel mapping framework also achieves 2–10% accuracy improvement as compared to the conventional approaches. Although the deep-learning-based approaches are computationally complex, the proposed super-resolution and sub-pixel mapping approaches are extensively parallelizable and offers significant reduction in execution time. In addition, the proposed inception-residual blocks improve the convergence as well as the accuracy when compared

to the simple convolution-deconvolution units. From the experiments over various datasets, the proposed frameworks are observed to outperform the prominent sub-pixel classification and super-resolution approaches. The data augmentation, proposed in this study, facilitates better generalization of the approach making it resilient to geometric and atmospheric distortions. The proposed spectral information based loss functions along with the perceptual similarity based ones improve the accuracy (avg. spectral similarity, SSIM and FSIM) by around 2–5% as compared to the mean squared error based approaches. Similarly, the proposed sub-pixel classification loss function reduces the number of misclassified pixels, by around 15–35%, in a computationally optimal manner. The sensitivity of the frameworks towards network parameter settings has been considerably reduced in our proposed models. The computational efficiency of the proposed models can be further improved by parallelizing the framework. Although advanced augmentation strategies improve generalization capability of the framework, spatial-spectral prior based refinements can be explored to further improve the universality. The current work was focused on off-line super-resolution and sub-pixel mapping of drone-derived datasets; however, the approach can be explored for on-the-fly analyses. In this regard, transfer learning based approaches, using the neighborhood video frames, can be adopted for improving the approach.

Acknowledgements

This research was supported by the Israel Ministry of Agriculture and Rural Development (Eugene Kandel Knowledge Centers) as part of the Precision agriculture – Development of systems to improve resources application in the field and partly (contract no. 235/16). Authors would like to thank the two anonymous reviewers and the editors whose considered efforts have significantly improved the quality of this work.

References

[1] A. Akyol, M. Gökmen, Super-resolution reconstruction of faces by enhanced global models of shape and texture, Pattern Recognit. 45 (12) (2012) 4103–4116.
[2] A. Antoniou, A. Storkey, and H. Edwards. Data augmentation generative adversarial networks. arXiv preprint, (2018) arXiv:1711.04340v3 [stat.ML].
[3] S. Anwar, C.P. Huynh, F. Porikli, Image deblurring with a class-specific prior, IEEE Transactions on. Pattern Analysis and Machine Mach. IntelligenceIntell (2018).
[4] P.V. Arun, K.M. Buddhiraju, Classification and clustering perspective towards spectral unmixing, in: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, 2016, pp. 6145–6148.
[5] P.V. Arun, K.M. Buddhiraju, A deep learning based spatial dependency modeling approach towards super-resolution, in: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, 2016, pp. 6533–6536.
[6] P.V. Arun, K.M. Buddhiraju, A. Porwal, Integration of contextual knowledge in unsupervised subpixel classification: semivariogram and pixel-affinity based approaches, IEEE Geosci. Remote Sens. Lett. 15 (2) (2018) 262–266.
[7] P.V. Arun, K.M. Buddhiraju, A. Porwal, CNN based sub-pixel mapping for hyperspectral images, Neurocomputing 311 (7) (2018) 51–64.
[8] J. Bergstra, D. Yamins, D. Cox, Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures, International Conference on Machine Learning, 2013 june 2013.
[9] M. Bevilacqua, A. Roumy, C. Guillemot, M.L. Alberi Morel, Neighbor embedding based single-image super-resolution using Semi-Nonnegative Matrix Factorization, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 1289–1292.
[10] D. Cai, K. Chen, Y. Qian, J. Kämäräinen, Convolutional low-resolution fine–grained classification, Pattern Recognit. Lett. (2017) 11.
[11] X. Chen, Z. Zhang, B. Wang, G. Hu, E.R. Hancock, Recovering variations in facial albedo from low resolution images, Pattern Recognit. 74 (2) (2018) 373–384.
[12] C. Dong, C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Analy. Mach. Intell. 38 (2) (2016) 295–307.
[13] W. Dong, F. Fu, G. Shi, X. Cao, J. Wu, G. Li, X. Li, Hyperspectral image super-resolution via non-negative structured sparse representation, IEEE Trans. Image Process. 25 (5) (2016) 2337–2352.
[14] A. Dosovitskiy, T. Brox, Generating images with perceptual similarity metrics based on deep networks, in: Proceedings of Advances in Neural Information Processing Systems-2016, 2016.
[15] A. Dosovitskiy, T. Brox, Inverting visual representations with convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 4829–4837.
[16] D. Eigen, D. Krishnan, R. Fergus, Restoring an image taken through a window covered with dirt or rain, in: International Conference on Computer Vision (ICCV), 2013, p. 2013.
[17] R. Feng, Y. Zhong, X. Xu, L. Zhang, Adaptive sparse subpixel mapping with a total variation model for remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 54 (5) (2016) 2855–2872.
[18] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. arXiv:1803.01229v1 [cs.CV].
[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of Advances in Neural Information Processing -2014, 2014.
[20] K. He, X. Zhang, S. Ren, and J. Sun, (2015). Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385.
[21] Z. He, L. Liu, S. Zhou, Y. Shen, Learning group-based sparse and low-rank representation for hyperspectral image classification, Pattern Recognit. 60 (12) (2016) 1041–1056.
[22] I. Herrmann, A. Pimstein, A. Karnieli, Y. Cohen, V. Alchanatis, D.J. bonfil, LAI assessment of wheat and potato crops by VENμS and Sentinel-2 bands, Remote Sens. Environ. 115 (2011) 2141–2151.
[23] H. Huang, H. He, X. Fan, J. Zhang, Super-resolution of human face image using canonical correlation analysis, Pattern Recognit. 43 (7) (2010) 2532–2543.
[24] J.B. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self-exemplars, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 5197–5206.
[25] S. Huang, J. Sun, Y. Yang, Y. Fang, P. Lin, Y. Que, Robust single-image super-resolution based on adaptive edge-preserving smoothing regularization, IEEE Trans. Image Process. 27 (6) (2018) 2650–2663.
[26] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 448–456.
[27] H. Irmak, G.B. Akar, S.E. Yuksel, A MAP-based approach for hyperspectral imagery super-resolution, IEEE Trans. Image Process. 27 (6) (2018) 2942–2951.
[28] J. Jebadurai, D.J. Peter, SK-SVR: Sigmoid kernel support vector regression based in-scale single image super-resolution, Pattern Recognit. Lett. 94 (7) (2017) 144–153.
[29] K. Jia, X. Wang, X. Tang, Image transformation based on learning dictionaries across image spaces, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2013) 367–380.
[30] H. Jin, G. Mountrakis, P. Li, A super-resolution mapping method using local indicator variograms, Int. J. Remote Sens. 33 (24) (2012) 7747–7773.
[31] L. Jin, Y. Zhang, S. Li, Integration-enhanced zhang neural network for real–time-varying matrix inversion in the presence of various kinds of noises, IEEE Trans. Neural Netw. Learn. Syst. 27 (12) (2016) 2615–2627.
[32] J. Johnson, A. Alahi, L. Fei-Fei, in: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. Computer Vision – ECCV 2016, Lecture Notes in Computer Science, 2016, p. 9906.
[33] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.
[34] N. Kumar, R. Verma, A. Sethi, Convolutional neural networks for wavelet domain super resolution, Pattern Recognit. Lett. 90 (4) (2017) 65–71.
[35] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Aitken, A. Te-jani, J. Totz, Z. Wang, and W. Shi, (2016). Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint, arXiv:1609.04802.
[36] K. Li, Y. Zhu, J. Yang, J. Jiang, Video super-resolution using an adaptive superpixel-guided auto-regressive model, Pattern Recognit. 51 (3) (2016) 59–71.
[37] Y. Li, W. Dong, X. Xie, G. Shi, J. Wu, X. Li, Image super-resolution with parametric sparse model learning, IEEE Trans. Image Process. 27 (9) (2018) 4638–4650.
[38] Y. Li, W. Xie, H. Li, Hyperspectral image reconstruction by deep convolutional neural network for classification, Pattern Recognit. 63 (3) (2017) 371–383.
[39] L. Liebel, M. Korner, Single-image super resolution for multispectral remote sensing data using convolutional neural networks, ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 3 (2016) 883–890.
[40] X. Liu, L. Chen, W. Wang, J. Zhao, Robust multi-frame super-resolution based on spatially weighted half-quadratic estimation and adaptive BTV regularization, IEEE Trans. Image Process. 27 (10) (2018) 4971–4986.
[41] A. Mahendran, A. Vedaldi, Understanding deep image representations by inverting them, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 5188–5196.
[42] Y. Makido, A. Shortridge, J.P. Messina, Assessing alternatives for modeling the spatial distribution of multiple land-cover classes at sub-pixel scales, Photogramm. Eng. Remote Sens. 73 (8) (2007) 935–943.
[43] D. Marmanis, J. Wegner, S. Galliani, K. Schindler, M. Datcu, U. Stilla, Semantic segmentation of aerial images with an ensemble of CNNs, ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 3 (2016) 473–480.
[44] M.T. McCann, K.H. Jin, M. Unser, Convolutional neural networks for inverse problems in imaging: a review, IEEE Signal Process. Mag. 34 (6) (2017) 85–95.
[45] K. Mertens, Sub-pixel mapping and sub-pixel sharpening using neural network predicted wavelet coefficients, Remote Sens. Environ. 91 (2) (2004) 225–236.
[46] K. Mertens, B. de Baets, L. Verbeke, R. de Wulf, A sub-pixel mapping algorithm based on sub-pixel/pixel spatial attraction models, Int. J. Remote Sens. 27 (15) (2006) 3293–3310.

[47] E. Pérez-Pellitero, J. Salvador, J. Ruiz-Hidalgo, B. Rosenhahn, PSyCo: manifold span reduction for super resolution, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 1837–1845.
[48] M.S.M. Sajjadi, B. Scholkopf, and M. Hirsch, (2016). EnhanceNet: single image super-resolution through automated texture synthesis. arXiv preprint, arXiv:1612.07919.
[49] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.R. Muller, Evaluating the visualization of what a deep neural network has learned, IEEE Trans. Neural Netw. Learn. Syst. 99 (8) (2016) 1–14.
[50] C.J. Schuler, H.C. Burger, S. Harmeling, B. Scholkopf, A machine learning approach for non-blind image deconvolution, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, p. 2013.
[51] K. Simonyan, and V. Zisserman, Very deep convolutional networks for large-scale image recognition. (2014). arXiv preprint arXiv:1409.1556.
[52] H. Su, N. Jiang, Y. Wu, J. Zhou, Single image super-resolution based on space structure learning, Pattern Recognit. Lett. 34 (16) (2013) 2094–2101.
[53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, (2015). Rethinking the inception architecture for computer vision. arXiv preprint, arXiv:1512.00567.
[54] C. Szegedy, S. Ioffe, and V. Vanhoucke, (2016). Inception-v4, inception-ResNet and the impact of residual connections on learning. arXiv preprint, arXiv:1602.07261v2.
[55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[56] A. Tatem, H. Lewis, P. Atkinson, M. Nixon, Increasing the spatial resolution of agricultural land cover maps using a Hopfield neural network, Int. J. Geograph. Inf. Sci. 17 (7) (2003) 647–672.
[57] X. Tong, X. Xu, A. Plaza, H. Xie, H. Pan, W. Cao, D. Lv, A new genetic method for subpixel mapping using hyperspectral images, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (9) (2016) 4480–4491.
[58] X. Tong, X. Zhang, J. Shan, H. Xie, M. Liu, Attraction-repulsion model-based subpixel mapping of multi-/hyperspectral imagery, IEEE Trans. Geosci. Remote Sens. 51 (5) (2013) 2799–2814.
[59] A. Villa, J. Chanussot, J.A. Benediktsson, C. Jutten, R. Dambreville, Unsupervised methods for the classification of hyperspectral images with low spatial resolution, Pattern Recognit. 46 (6) (2013) 1556–1568.
[60] C. Wang, C. Xu, C. Wang, D. Tao, Perceptual adversarial networks for image–to-image transformation, IEEE Trans. Image Process. 27 (8) (2018) 4066–4079.
[61] L. Wang, Z. Huang, Y. Gong, C. Pan, Ensemble based deep networks for image super-resolution, Pattern Recognit. 68 (2017) 191–198.
[62] Q. Wang, P.M. Atkinson, W. Shi, Indicator Co-kriging-based subpixel mapping without prior spatial structure information, IEEE Trans. Geosci. Remote Sens. 53 (1) (2015) 309–323.
[63] Q. Wang, W. Shi, P.M. Atkinson, Z. Li, Land cover change detection at subpixel resolution with a hopfield neural network, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8 (3) (2015) 1339–1352.
[64] Q. Wang, Y. Tao, C. Wang, F. Dong, H. Lin, G. Clapworthy, Volume upscaling using local self-examples for high quality volume visualization, in: International Conference on Computer-Aided Design and Computer Graphics, Guangzhou, 2013, pp. 298–305.
[65] Q. Wang, L. Wang, D. Liu, Integration of spatial attractions between and within pixels for sub-pixel mapping, J. Syst. Eng. Electron. 23 (2) (2012) 293–303.
[66] X. Wei, Y. Li, H. Shen, W. Xiang, Y.L. Murphey, Joint learning sparsifying linear transformation for low-resolution image synthesis and recognition, Pattern Recognit. 66 (2) (2017) 412–424.
[67] Z. Wei, B. Xiaofeng, H. Fang, W. Jun, A.A. Mongi, Fast image super-resolution algorithm based on multi-resolution dictionary learning and sparse representation, J. Syst. Eng. Electron. 29 (3) (2018) 471–482.
[68] B. Xu, N. Wang, T. Chen, M. Li, (2015). Empirical evaluation of rectified activations in convolutional network, arXiv preprint, arXiv:1505.00853 [cs.LG].
[69] L. Xu, J.S. Ren, C. Liu, J. Jia, Deep convolutional neural network for image deconvolution, Neural Information processing Systems – 2014, 2014.
[70] X. Xu, Y. Zhong, L. Zhang, Adaptive subpixel mapping based on a multiagent system for remote-sensing imagery, IEEE Trans. Geosci. Remote Sens. 52 (2) (2014) 787–804.
[71] J. Yang, Z. Wang, Z. Lin, S. Cohen, T. Huang, Coupled dictionary training for image super-resolution, IEEE Trans. Image Process. 21 (8) (2012) 3467–3478.
[72] S. Yang, M. Wang, Y. Sun, F. Sun, L. Jiao, Compressive sampling based single-image super-resolution reconstruction by dual-sparsity and non-local similarity regularizer, Pattern Recognit. Lett. 33 (9) (2012) 1049–1059.
[73] C. Yi, Y.Q. Zhao, J.C.W. Chan, Hyperspectral image super-resolution based on spatial and spectral correlation fusion, IEEE Trans. Geosci. Remote Sens. 56 (7) (2018) 4165–4177.
[74] F. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, StackGAN++: realistic image synthesis with stacked generative adversarial networks, IEEE Trans. Pattern Anal. Mach. Intell. 2018 (7) (2018) 1.
[75] Y. Zhang, D. Guo, Z. Li, Common nature of learning between back-propagation and hopfield-type neural networks for generalized matrix inversion with simplified models, IEEE Trans. Neural Netw. Learn. Syst. 24 (4) (2013) 579–592.
[76] J. Zhao, Y. Zhong, Y. Wu, L. Zhang, H. Shu, Sub-pixel mapping based on conditional random fields for hyperspectral remote sensing imagery, IEEE J. Sel. Top. Signal Process. 9 (6) (2015) 1049–1060.
[77] Y. Zhong, L. Zhang, Sub-pixel mapping based on artificial immune systems for remote sensing imagery, Pattern Recognit. 46 (11) (2013) 2902–2926.
[78] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2921–2929.

Arun P.V. is a second year PhD student at IIT Bombay and visiting scholar at Ben-Gurion University of Negev. His areas of expertise are hyperspectral image analysis, deep learning, and remote sensing.

Ittai Herrmann received his PhD at Ben-Gurion University of The Negev, Sede Boker Campus, Israel in 2012. He explored avian eggs spectroscopy in an Israeli startup company. He returned to the Remote Sensing Laboratory, Ben-Gurion University of the Negev as a postdoctoral researcher. Starting 2016 he is a postdoctoral researcher in The Townsend Lab, Department of Forest and Wildlife Ecology in conjunction with The CoolBean program, extension, Department of Agronomy, both in University of Wisconsin - Madison, Madison, USA. His main interest is in multi- and hyperspectral data analysis of vegetation focusing in precision agriculture.

Krishna Mohan Buddhiraju is the chair professor and Head of department at Centre for Studies in Resources Engineering, IIT Bombay. He is also the coordinator of ISRO-IITB space technology cell. His areas of expertise include machine learning, remote sensing and computer vision.

Arnon Karnieli received the Ph.D. degree from the University of Arizona, in 1988. Since then, he has been the Head of the Remote Sensing Laboratory, Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Sede Boker Campus, Israel. His main research is focused on processing of spaceborne, airborne, and ground spectroscopic data of drylands with respect to desertification and climate change processes. In this regard, his study's applications cover dryland ecosystems and agriculture, and to a lesser extent dust and coastal water. He is the Israeli Principle Investigator of the forthcoming Vegetation and Environmental New Micro Spacecraft (VENμS) mission. Prof. Karnieli has published more than 170 papers in peer-reviewed journals. Beside Israel, his world-wide activities were conducted in Central Asia, Mongolia, China, and West Africa.