MU-Net: Deep Learning-based Thermal IR Image Estimation from RGB Image

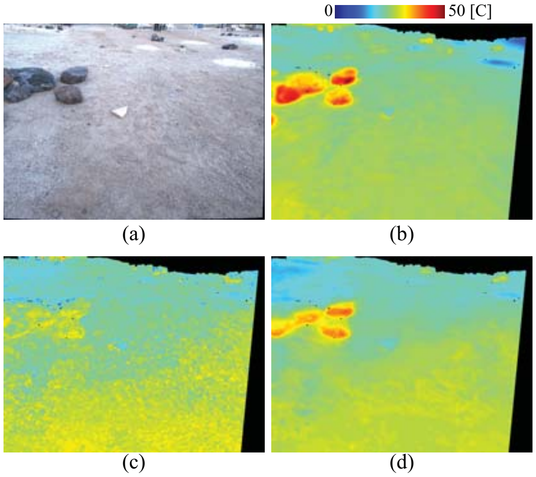
Yumi Iwashita[1], Kazuto Nakashima[2], Sir Rafol[1], Adrian Stoica[1], Ryo Kurazume[2]
[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
[2]Kyushu Universeity, Fukuoka, Japan
Yumi.Iwashita, Sir.B.Rafol, Adrian.Stoica@jpl.nasa.gov
k_nakashima, kurazume@irvs.ait.kyushu-u.ac.jp

Abstract

Terrain imagery collected by satellite remote sensing or by rover on-board sensors is the primary source for terrain classification used in determining terrain traversibility and mission plans for planetary rovers. Mapping models between RGB and IR for terrain classes are learned from real RGB and IR data examples in the same or similar terrain. This paper adds a new class of deep learning architectures called MU-Net (Multiple U-Net) and shows its efficiency in deriving better RGB-to-IR mapping models, improving over past work the estimation of thermal IR images from incoming RGB images and learned RGB-IR mappings. [1]

1. Introduction

Terrain classification is one of the key components for autonomous navigation for Mars rovers. A terrain classification system that uses both RGB and thermal infrared (IR) images to improve the performance of terrain classification compared to using RGB was proposed in [1]. However, while future rovers may have IR cameras, neither the Mars Science Laboratory (MSL) nor Mars 2020 [2] have IR cameras, relying solely on RGB cameras. A possible way to circumvent this absence is by using, instead of real IR, estimates of IR learned from examples of images seen both in RGB and IR. Learning from these examples was shown with a new deep learning technique demonstrated in [3]. Figures 1 (a) and (b) show an example of images with RGB and IR cameras respectively, and Fig. 1 (c) shows an estimated IR image from Fig. 1 (a) based on a UNet-based method proposed in [3]. A new learning architecture set introduced in this paper leads to better learned models and improved IR estimation as shown in Figure 1 (d) (derived from learned mapping and input RGB as in for the Fig. 1 (a).



(a)

(b)

(c)

(d)

Figure 1. (a) RGB image, (b) ground truth thermal IR image, (c) estimated thermal IR [3], and (d) estimated thermal IR with the proposed network. (a) and (b) were captured at 5 pm on Nov 17th.

The experimental results, albeit limited so far, show that, at least for the limited and constrained conditions of terrains similar to planetary surfaces with relatively low diversity of components (various rocks and sands), the estimation of the IR domain from the RGB domain appears possible (and could be very useful for autonomous driving and for science).

One possible explanation of this first set of encouraging results of estimation of IR from RGB, in given conditions, could be related to the balance of energy exchange at the imaged object. Over a short duration of time, the energy emitted by a surface object is approximately equivalent to the absorbed energy. In daytime, incoming energy is from the sun with known spectrum, some of it reflected, which includes the part observable in RGB/visual, and some is absorbed, which is about the same as reflected (assuming opaque objects). The emitted energy depends on the emissivity and temperature of material. Emissivity depends on the type of material and wavelength. In daytime, temperature is determined mainly by how much energy hits the surface normally (i.e., angle of the surface to the Sun) and thermal inertia (which shows a measure of material's ability to resist a change in temperature). In the case that slope angle difference among different types of terrains is small, the temperature of each terrain is dominated by its thermal inertia. Terrain type, through thermal inertia, thus influences temperature of the terrain and hence provides thermal IR information.

While the earlier work indicated the feasibility of estimating IR from RGB, the performance was moderate [3]. No possible theoretical interpretation was attempted and no integration of terrain type information in the model was attempted.

This paper first attempts a theoretical explanation that may support the results of estimation of IR images from RGB images. A new set of deep learning architectures taking into account the terrain type information is introduced. The new set of architectures, named MU-Net (from Multiple U-Net), is based on U-Net [4], which is popularly used in medical image segmentation [5] and also was used by the winner of a satellite image segmentation competition (Kaggle competition [6]). MU-Net is designed to estimate terrain type information in addition to thermal IR information. This allows us to both ***implicitly*** and ***explicitly*** include terrain type information to estimate thermal IR information, which results in improvement of the estimation of thermal IR images from a single sensor input (RGB camera).

There is a resemblance between the estimation of IR from RGB (for which we are not aware of any other work except [3]) and colorization of gray scale images, for which a body of work exists [7] [8] [9]. In general these methods require estimation of chrominance, since the luminance is given in the grayscale images. Iizuka et al. proposed a deep convolutional neural networks (deep CNN) to directly estimate chrominance values in gray-scale images [11]. Larsson et al. [12] and Zhang et al. [13] initialized their networks with pre-trained networks. Limmer et al. proposed a CNN-based method to colorize near IR images, which requires estimation of chrominance and luminance [10].

Deep neural network learning requires huge datasets; e.g. in just cited [10] almost 38,495 image pairs are used. If not enough training data is available one can work with pre-trained parameters with public datasets, such as ImageNet [14]. However, since public datasets usually include images with huge inter-class variations, such as cars, human, balls, etc, the pre-trained parameters do not efficiently describe features of terrain types, where inter-class variations are much smaller than in public datasets. On the other hand, training using smaller datasets may be feasible via U-net.

2. Methodology

This section first provides a possible theoretical explanation of why one can potentially estimate IR images from RGB images. The U-net method to estimate thermal IR information [3] is then explained, followed by the proposed MU-Net.

RGB cameras respond to wavelengths from about 390 to 700 [nm] while thermal cameras respond to different wavelengths, such as 7-14 [μm] for long-wave IR. In general information in one spectral domain cannot be definitely determined by information in a different spectral domain. However, under certain assumptions, specifically, (i) the material is opaque, (ii) the event happens instantaneously, and (iii) the radiating source is only the Sun, we can estimate a representation of IR domain from RGB domain.

In the case of an opaque material, incident energy I is defined with reflection energy R and absorbed energy A as $I = R + A$. For the material to stay in equilibrium, absorbed energy A should be equal to emission energy E (E=A). (More correctly, we refer the energy for a short duration of time, so effectively the power). Thus, the first equation becomes $I = R + E$. Here, we can assume the RGB images capture reflection R from the material and the IR images capture emission E from it. The emission energy E observed in IR images, is now defined as $E = I - R$. A main parameter of the incident I is the angle to the Sun, and the parameter can be calibrated based on the Sun's angle and geometric information. In a simplifying assumption one can assume the angle to the Sun is uniform ($i.e.$, the incident I is constant in the whole area) (meaning also the angle of the receiving/reflecting surface is the same in all area of interest). These assumptions are too strong and do not hold except for rare cases, so there is no quality However, some relationships exist and hence estimates limited, but potentially useful.

The architecture of U-Net [4] is shown in Fig. 2 and consists of a contracting path (left) and an expansive path (right). Each path has repeated units. The unit on the contracting path (contracting unit), as shown with light blue rectangles, consists of two 3 × 3 convolutions, each followed by a rectified linear unit (ReLU) and 2 × 2 max-pooling. There are two different units on the expansive path. The first one (expansive unit 1) as shown with red rectangles consists of two 3 × 3 convolutions, each followed by ReLU, 2 × 2 deconvolution, and concatenation of outputs from both deconvolution layer and convolution layer from

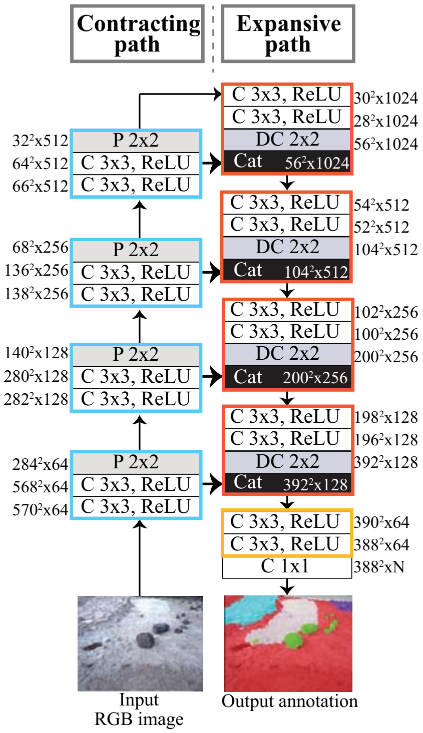


Figure 2. U-Net architecture. "Cat", "C", "ReLU", "P", and "DC" mean "concatenate", "convolution", "rectified linear unit", "pooling", and "deconvolution", respectively. Relatively thick arrows between "Cat" and "C" include "bilinear up-sample". "N" at the final layer show the number of classes. Light blue rectangles show units of the contracting path (contracting units). Red and orange rectangles show two different units of expansive paths (expansive units 1 and 2).

the contracting path. Another one (expansive unit 2) is the last unit of the expansive path, as shown with an orange rectangle, has two 3×3 convolutions, each followed by a rectified linear unit (ReLU). Here, in the expansive unit 1, bilinear up-sample is applied to output of the convolution layer from the contracting path. This concatenation layer is one of key ideas in U-Net, which enables training of the network with a small number of datasets. At the final layer, a 1×1 convolution is applied to map 64 channel information at each pixel to the number of classes (N).

The loss function $\mathcal{L}_{CE}$ of all architectures is defined as a pixel-wise soft-max over the final map, followed by the cross-entropy loss function, as defined as follows.

$$\mathcal{L}_{CE} = -\frac{1}{|S|} \sum_{i \in S} \sum_{j=1}^{N} y_{ij} \log p_{ij}, \tag{1}$$

where N, $|S|$, y_{ij}, p_{ij} are the number of classes, the total number of pixels over images S, ground-truth distribution at each pixel, and outputted probability distribution at each pixel, respectively. The loss function is minimized by a stochastic gradient descent method.

To synthesize thermal IR images from RGB images, we replaced the output annotation in Fig. 2 with a thermal IR image. We used a mean squared error (MSE) $\mathcal{L}_{MSE}$ as a loss function, which is defined as

$$\mathcal{L}_{MSE} = \frac{1}{|S| \times C} \sum_{i \in S} \sum_{j=1}^{C} (a_{ij} - b_{ij})^2, \tag{2}$$

where C is the number of channels, and a_{ij} and b_{ij} are the thermal value at each pixel (i, j) of a ground-truth thermal image a and an output thermal image b, respectively.

This model directly trains the network with RGB images and does not take into account terrain-type information, which has a potential to improve the estimation of thermal IR images.

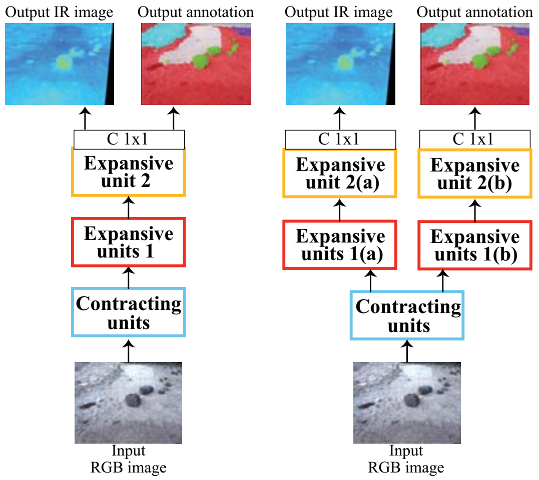
To take into account terrain-type information, we propose MU-Net (Multiple U-Net). The proposed MU-Net has two categories: (i) MU-Net1, implicitly including terrain-type information into the model and (ii) MU-Net2, explicitly including terrain type information into the model.

MU-Net1 is designed to output both IR thermal and annotation images, so that the trained model includes both IR and annotation information. MU-Net1 has two architectures as shown in Fig. 3. The first one, MU-Net1-a is the same architecture with U-Net, but MU-Net1-a has two outputs of annotation and IR thermal images. Thus, in this model we have two loss functions, $\mathcal{L}_{CE}$ and $\mathcal{L}_{MSE}$, and have combined them as a weighted loss function as

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{MSE}. \tag{3}$$

Here, λ is empirically assigned the value of 200. The model of MU-Net1-a is trained in a way that both loss functions of IR thermal and annotation images are minimized. The second architecture is MU-Net1-b as shown in Fig. 3 (b), and this architecture has expansive units for each IR thermal and annotation images. This is because of the following reason. IR thermal and annotation images show fundamentally different images; thus MU-Net1-a may not be able to model these two images in the common expansive units. In MU-Net1-b also uses the weighted loss function Eq. 3.

MU-Net2 is designed to explicitly include terrain type information into the architectures (Figs. 4 and 5). MU-Net2 has an independent architecture for annotation images as shown in dotted gray rectangles in Figs. 4 and 5. The model for annotation images is trained first, and trained parameters are then copied to architectures for IR thermal images. We also have two different architectures for MU-Net2. The first one, MU-Net2-a (Fig .4), has two different contracting units (a) and (b), and contracting units (b) are copies from the model for annotation images. Outputs from both contracting units are concatenated and used as input to



(a) (b)

Figure 3. (a) MU-Net1-a and (b) MU-Net1-b.

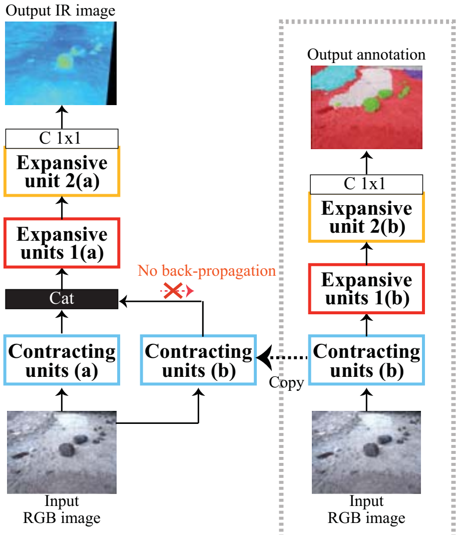


Figure 4. MU-Net2-a.

the expansive units 1(a). There is no back-propagation to contracting units (b). The second architecture, MU-Net2-b (Fig. 5), copies parameters of contracting units (b), and expansive units 1(b) and 2(b) of the model for annotation images in the dotted gray rectangle in Fig. 5. The output of expansive units 2(b) is downscaled with a pixel-shuffle technique, and it is concatenated with the output of contracting units (a). Finally, it is used as input to the expansive units 1(a). For both MU-Net2-a and MU-Net2-b, we use only $\mathcal{L}_{MSE}$ as the loss function of IR images.

3. Experiments

In this section, we first explain a dataset which includes visible and thermal images, followed by experimental results with the dataset.

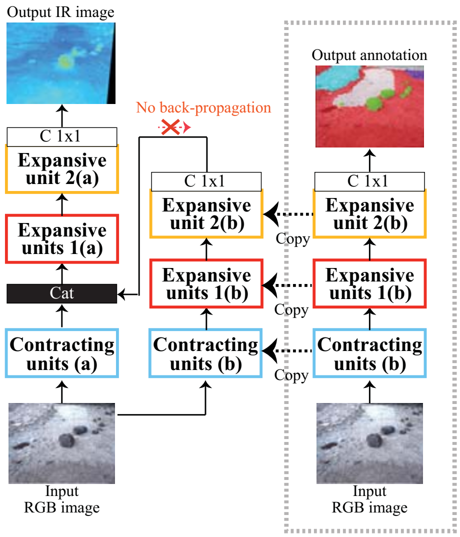
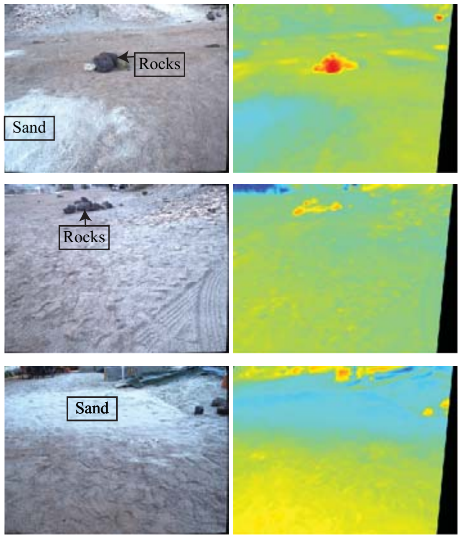


Figure 5. MU-Net2-b.



(a) (b)

Figure 6. (a) Examples of RGB images and (b) IR image corresponding to (a).

We used the same dataset in [3] for training and we added more dataset for performance evaluation. The images were collected at an area (the JPL Mars Yard) with a RGB camera (FLIR Grasshopper 5M) and a thermal camera (FLIR AX65) from 10am to 5pm on Nov 17th, 2017. RGB and IR images were collected every 1 hour, with 52 images images collected each time by changing the position of the cameras (totalling 416 image pairs over the 8 hours). Figure 6 shows examples of captured image pairs. Sandy areas tend to show lower temperature due to the fact of its lower thermal inertia. On the other hand, rocky areas show higher temperatures since they have higher thermal inertia. Since the visible and thermal images were taken by different cameras,

Table 1. Mean absolute error (MAE) of thermal IR images estimated from RGB images taken at 5 pm with normalized estimated IR and normalized ground truth IR images (i.e. E'_e and E'_g). Comparison of 5 methodologies ((a) U-Net [3], (b) MU-Net1-a, (c) MU-Net1-b, (d) MU-Net2-a, and (e) MU-Net2-b).

	(a) U-Net [3]	(b) MU-Net1-a	(c) MU-Net1-b	(d) MU-Net2-a	(e) MU-Net2-b
MAE	0.69	0.61	**0.59**	0.70	0.71

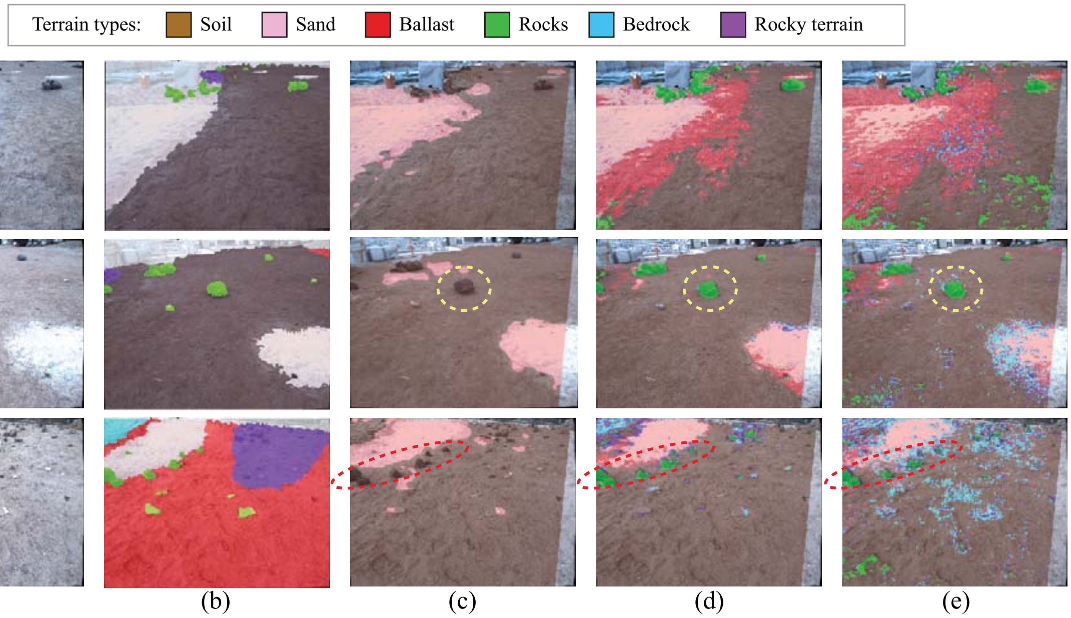


Figure 7. (a) Examples of RGB images in test dataset, (b) manually annotated images corresponding to (a), (c) estimated terrain types by MU-Net1-a, (d) estimated terrain types by MU-Net1-b, and (e) estimated terrain type by MU-Net2.

a registration process between cameras is necessary. After we removed distortion with estimated camera inner parameters, we applied an affine transformation with an estimated homography matrix.

In the following experiments, we used 50% of RGB and IR images at 5 pm as gallery data and 25 % at 5 pm for evaluation to determine parameters. There are two settings for performance evaluation: (i) the rest 25% at 5 pm for the test, and (ii) images taken at every 1 hour from 10 am to 5 pm for test. As for the terrain classification, we categorized the area into 6 terrain types (soil, sand, rocks, bedrocks, rocky terrain, and ballast). The data size of each terrain type is not balanced, so we introduced weights to the MSE and cross-entropy losses. Here, the assumption is that each terrain has a unique temperature on IR images, and we ignore other factors which change temperature, such as shades and slopes. The weight of each terrain type is defined as the squared root of the ratio of number of pixels in the training dataset. Annotation images as shown in Fig. 7 (a) are set in advance manually.

In the first experiments, we applied the proposed MU-Net1, which implicitly utilizes terrain type information into the model, and the proposed MU-Net2, which explicitly utilizes terrain type information into the model, to the dataset. Figures 7 (a) $\sim$ (e) show examples of captured images, manually annotated images, estimated annotation images by MU-Net1-a, those by MU-Net1-b, and those by MU-Net2, respectively. Estimated terrain types by MU-Net1-a tends to include more false positives than those by MU-Net1-b. For example rocks estimated by MU-Net1-a (Fig. 7 (c)) are misclassified as soil. This suggests that expansive units for each thermal IR images and annotation images works effectively. The estimated terrain types by MU-Net2 in Fig. 7 (e) include more false positives than MU-Net1 (Figs 7 (c) and (d)). The bottom figures in Figs. 7 (c) $\sim$ (e) show that ballast area (red area) is misclassified as soil area (brown area), but this area is a mixed area of soil and ballast area, which is difficult even for people to classify it.

As for the IR images estimated from RGB images, first we show quantitative evaluations as shown in Table 1. As we mentioned in section 2, the estimated IR values are scaled values of actual values. Since we cannot directly compare ground truth IR values $E_g(x, y)$ and estimated IR values $E_e(x, y)$, we normalize the values based on standard deviation and mean as $E'_g(x, y)$ = ($E_g(x, y)$ - μ_g) / σ_g and $E'_e(x, y)$ = ($E_e(x, y)$ - μ_e) / σ_e, where μ_g, σ_g, μ_e, and σ_e

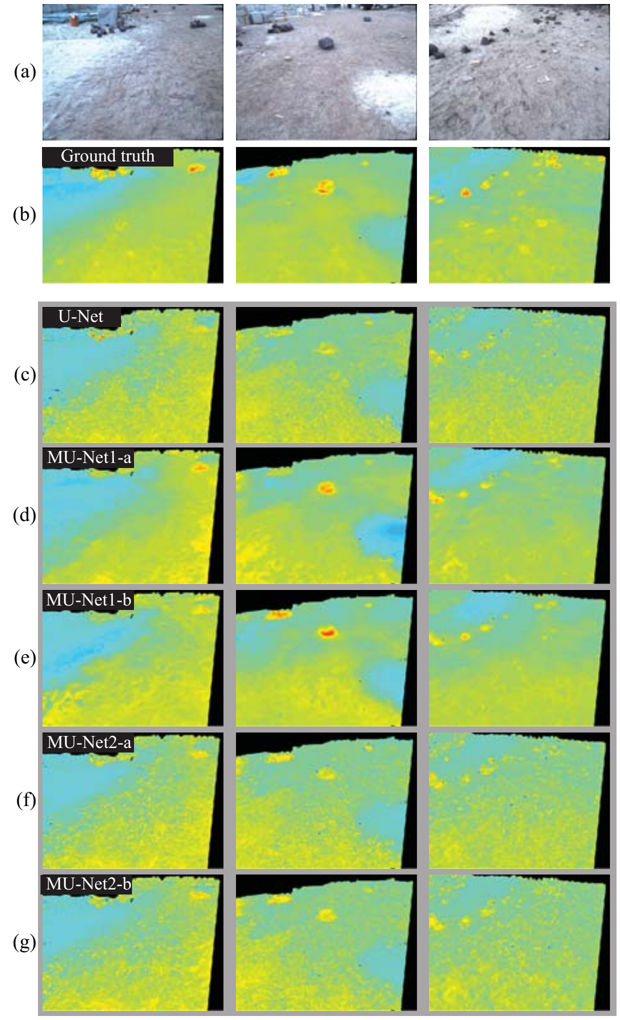


Figure 8. (a) Examples of RGB images in test dataset, (b) ground truth thermal IR images corresponding to (a), (c) estimated thermal IR images by U-Net [3], (d) estimated thermal IR images by MU-Net1-a, and (e) estimated thermal IR images by MU-Net1-b, (f) estimated thermal IR images by MU-Net2-a, and (g) estimated thermal IR images by MU-Net2-b.

are mean and standard deviation of the ground truth IR and the estimated IR images. From the normalized ground truth IR images and the normalized estimated IR images, a mean absolute error (MAE) is calculated for each approach. MU-Net1-b shows the smallest error among the all models.

We also visualized examples of ground truth IR images and corresponding estimated IR images as shown in Fig. 8 using the normalized images, as $E''_g(x,y) = (E'_g(x,y) * \sigma_g + 128)$ and $E''_e(x,y) = (E'_e(x,y) * \sigma_g + 128)$. Figures 8 (a), (b), (d), and (e) show examples of captured images, ground truth IR thermal images corresponding to (a), estimated IR thermal images by MU-Net1-a, and those by MU-Net1-b, respectively. We also compared the proposed MU-Net1 with [3], whose results are shown in Fig. 8 (c). Results by MU-Net1-b show smoother results than those by [3]. These results also suggest that the performance of MU-Net1-b is better than that of U-Net and MU-Net1-a. Figures 8 (f) and (g) show estimated thermal IR images by MU-Net2-a and MU-Net2-b, and these results show more false positives than the results of MU-Net1-b. MAE of MU-Net1-a based on E''_g and E''_e is 2.30 degree.

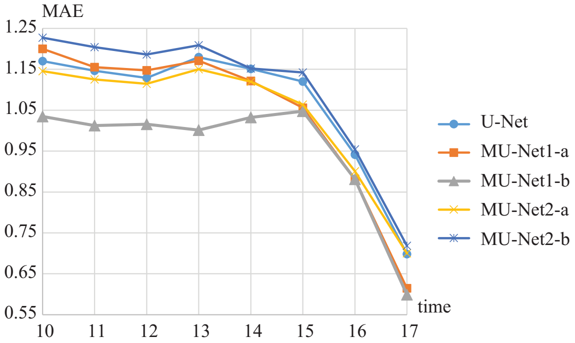


Figure 9. MAEs (mean absolute error) of MU-Net1-a, MU-Net1-b, MU-Net2-a, and MU-Net2-b. Models are trained with images at 5 pm and tested with images from 10 am to 5 pm. MU-Net1-b shows the smallest MAE.

From the above results, MU-Net1-a and MU-Net1-b perform better than MU-Net2-a and MU-Net2-b. One of the reasons why MU-Net1 is better is as follows. In MU-Net2, the model for the annotation images is trained with only annotation images. On the other hand, MU-Net1 trains the network with both annotation and thermal IR images. The use of IR images in MU-Net1 gives additional constrains which improve the performance of the classification of annotation images. This results in improving the estimation of thermal IR images in MU-Net1.

In our next experiments, we used images taken at every 1 hour from 10 am to 5 pm as a probe dataset, to see if the models trained with images at 5 pm are robust in time variations. As for the experiment setting of images at 5 pm, we used the same setting as with the previous section (i.e., no overlap among gallery, evaluation, and test dataset). Figure 9 shows MAEs of MU-Net1-a, MU-Net1-b, MU-Net2-a, and MU-Net2-b from 10 am to 5 pm. These results show that MU-Net1-b is the most robust architecture among the five architectures.

Figure 10 (a) shows captured RGB images from 10 am to 4 pm, and Figs. 10 (b) and (c) shows ground-truth IR images corresponding to (a) and estimated thermal IR images by MU-Net1-b. These images show that overall temperature characteristics are predicted by the proposed method, but we can see differences between ground-truth IR images and estimated IR images, because the proposed method does not take into account shadows, angle of the Sun, geological information, etc. These are left for a future work.

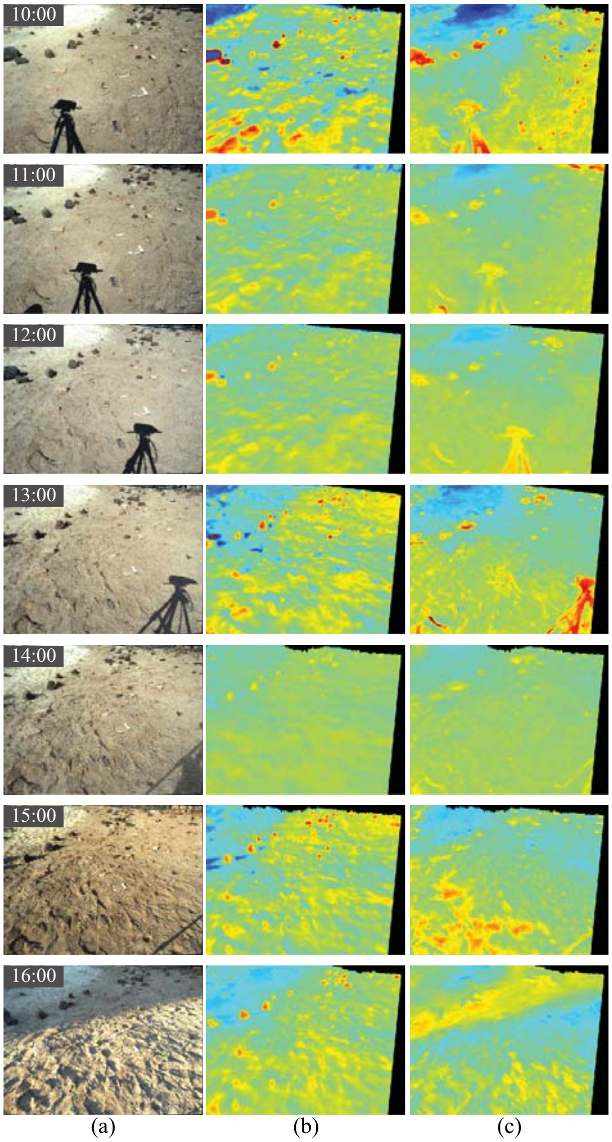


Figure 10. (a) Examples of RGB images from 10 am to 4 pm, (b) ground truth IR images corresponding to (a), (c) estimated thermal IR images by MU-Net1-b.

4. Conclusion

In this paper we proposed the use of terrain type information to estimate thermal IR images from RGB images. We introduced four deep learning architectures called MU-Net1-a, MU-Net1-b, MU-Net2-a, and MU-Net2-b. MU-Net1-b showed the best performance, since it takes advantages of using annotation images as constraints in addition to thermal IR images to train the model. There are many parameters to determine the temperature of terrain surface, such as thermal inertia, direction to the Sun, geological condition, etc. Future work will include these parameters in the model.

5. Acknowledgment

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

[1] Y. Iwashita, K. Nakashima, A. Stoica, R. Kurazume, TU-Net and TDeepLab: Deep Learning-based Terrain Classification Robust to Illumination Changes, Combining Visible and Thermal Imagery, IEEE Int. Conf. on Multimedia Information Processing and Retrieval 2019, accepted.

[2] Mars Science Laboratory (MSL), https://mars.nasa.gov/msl/

[3] Y. Iwashita, A. Stoica, K. Nakashima, Virtual sensors determined through machine learning, World Automation Congress (WAC) 2018.

[4] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, MICCAI 2015.

[5] O. Cicek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, MICCAI 2016.

[6] Dstl Satellite Imagery Feature Detection, https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection

[7] Z. Cheng, Q. Yang, and B. Sheng, Deep colorization, Proc. International Conference on Computer Vision ,2015.

[8] G. Patterson and J. Hays, Sun attribute database: Discovering, annotating, and recognizing scene attributes, Proc. Conference on Computer Vision and Pattern Recognition, 2012.

[9] E. Tola, V. Lepetit, and P. Fua, A fast local descriptor for dense matching, Proc. Conference on Computer Vision and Pattern Recognition, 2008.

[10] M. Limmer, and H. Lensch Infrared Colorization Using Deep Convolutional Neural Networks, IEEE Int. Conf. on Machine Learning and Applications (ICMLA), 2016.

[11] S. Iizuka, E. Simo-Serra, and H. Ishikawa, Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification, Proc. ACM SIGGRAPH, vol. 35, no. 4, 2016.

[12] G. Larsson, M. Maire, and G. Shakhnarovich, Learning representations for automatic colorization, arXiv:1603.06668, 2016.

[13] R. Zhang, P. Isola, and A. A. Efros, Colorful image colorization, arXiv:1603.08511, 2016.

[14] J. Den, W. Dong, R. Socher, and F. Li, ImageNet: a Large-Scale Hierarchical Image Database, CVPR 2009.