Supervised conversion from Landsat-8 images to Sentinel-2 images with deep learning

Sani M. Isa a, Suharjito a, Gede Putera Kusuma a and Tjeng Wawan Cenggoro b,c

aComputer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia; bComputer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia; cBioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia

ABSTRACT

In a specific remote sensing study design, the utilization of images from a particular satellite is necessary. However, the images might be unavailable in a certain time range. Therefore, a conversion method from available remote sensing images at the time range is required. In this paper, we proposed machine learning models that are capable to convert Landsat-8 images to Sentinel-2 images. The models are inspired by the advancement of super-resolution model based on Deep learning. The result of this study shows that the proposed models can predict Sentinel-2 images which are quantitatively and qualitatively similar to the original images.

Introduction

The availability of images from a suitable satellite is an essential factor in designing a study in remote sensing. However, the options for satellite are constrained with the launching time, the terminating time, and the time resolution of the satellite. Suppose that we want to conduct a study using satellite images captured in 2013. In this case, Sentinel-2 images are not available because the satellite was launched on June 23, 2015. As an alternative, Landsat-8 images can be used. However, in some study designs, Sentinel-2 images are preferred due to its higher spatial resolution. Therefore, there is a demand for a model capable to convert images between different satellites. Unfortunately, the development of a conversion model is not well studied. The most similar models that can be found are the model developed in the remote sensing studies for image fusion, pan-sharpening, and image super-resolution.

Image fusion studies aim to combine multiple images from different satellites into new images that possess favored characteristics from the original images. An example of the studies is the harmonization Landsat-8 – Sentinel-2 by Claverie et al. (2018). In the study, Sentinel-2 images and Landsat-8 images were fused in a virtual constellation model to form three images: (1) S10, which is a Sentinel-2-like image; (2) S30, which is an image with combined characteristics of Sentinel-2 and Landsat-8; and (3) L30, which is a Landsat-8-like image. Other examples are the spatio-temporal image fusion studies, whose objective is to fuse a low spatial – high temporal resolution image with a high spatial – low temporal resolution image to produce a high spatial – high temporal resolution image. The popular methods for spatio-temporal image fusion, according to Belgiu and Stein (2019), are STARFM (Gao et al., 2006), ESTARFM (Zhu et al., 2010), and STAARCH (Hilker et al., 2009). Following the trend in computer vision research, the recent studies in this topic are starting to adopt deep learning approaches. Ma et al. (2019) identified three recent image fusion studies that used deep learning: the study by Palsson et al. (2017), Dian et al. (2018), and Yang et al. (2018).

On the other hand, pan-sharpening studies objective is to improve the spatial resolution of a single image with the help of a high spatial resolution band named the panchromatic band. A pan-sharpening studies can also be categorized as an image fusion study (Ghamisi et al., 2019), in the view that they fuse low spatial resolution bands with panchromatic bands to output high spatial resolution bands. Similar to image fusion research, the shift to deep learning approaches is apparent in this topic. As reported by Ma et al. (2019), at least seven pan-sharpening studies have utilized deep learning (Huang et al., 2015; Masi et al., 2016; Scarpa et al., 2018; Shao & Cai, 2018; Wei et al., 2017; Xing et al., 2018; Yuan et al., 2018). Furthermore, we identified two additional deep-learning-based pan-sharpening studies (Masi et al., 2017; Yang et al., 2017).

Similar to pan-sharpening studies, the goal of super-resolution studies is to enhance the spatial resolution of an image. However, panchromatic band is not utilized in a super-resolution model. These studies are largely influenced by the super-resolution models in computer vision research, which are mostly based on the Fully
Convolutional Network (FCN) (Long et al., 2015). The earliest super-resolution study in remote sensing was done by Collins et al. (2017). In the work, a SRCNN-based network (Dong et al., 2014) was used to enhance the spatial resolution of Resourcesat-1 and -2 images. A year later, Pouliot et al. (2018) developed a modified SRCNN called as DCR_SRCNN to enhance the red, Near Infrared (NIR), and Shortwave Infrared (SWIR) bands of Landsat-8 by using Sentinel-2 red, NIR, and SWIR bands. The work by Pouliot et al. can also be considered as an image fusion study because of the use of Sentinel-2 as an auxiliary input. In the same year, Lanaras et al. (2018) developed a deep-learning-based super-resolution model specifically for Sentinel-2 images. The super-resolution technique is also proved to be capable for enhancing the resolution of hyperspectral aerial images (W. Liu & Lee, 2019; Ran et al., 2019). Not only to enhance the spatial resolution of aerial images, but the super-resolution technique can also be applied to enhance the spatial resolution of land cover maps. Jia et al. (2019) showed that CNN can predict a finer land cover map given a coarse remote sensing image.

Despite the problem similarity, it is unfortunate that the model developed for image fusion, pan-sharpening, and image super-resolution cannot be directly used to solve image conversion problem. For instance, in Landsat-8 to Sentinel-2 conversion, the model should be able to generate a Sentinel-2 image with the input of only a Landsat-8 image, as illustrated in Figure 1. It is different from the image fusion model, which combines two images from different satellites to form a new image as seen in Figure 2. Similarly, it also differs from the pan-sharpening model due to the use of a panchromatic band, as depicted in Figure 3. Among the three models, the image super-resolution model is the most similar model to the conversion model. However, the super-resolution model outputs an image from the same satellite with an improved spatial resolution instead of an image from another satellite, as visualized in Figure 4.

The uniqueness of the conversion model approach enables it to perform tasks that are not possible for the image fusion, pan-sharpening, and image super-resolution model. For instance, a conversion model of Landsat-8 to Sentinel-2 can predict a Sentinel-2 image dated between February 11, 2013 and June 23,
2015, a period when Landsat-8 has been orbited before the launch of Sentinel-2.

To develop the conversion model, we propose to use deep learning approach, which showed superior performance in a wide range of remote sensing research sub-domains (Gu et al., 2019). The sub-domains are not only limited to the image fusion, pan-sharpening, and remote sensing image super-resolution as discussed before, but also other sub-domains with different characteristics such as scene classification (Chaib et al., 2017; Cheng et al., 2018, 2016; F. Zhang et al., 2016), object detection in remote sensing image (Cao et al., 2016; Q. Liu et al., 2018; Ren et al., 2018; W. C. Zhang & Chen, 2012), and scene retrieval (Han et al., 2020; Li et al., 2020, 2018).

This study was focused to develop only Landsat-8 image to Sentinel-2 image conversion models. However, the model should be able to generalize to images from any satellites. The contributions of this study are:

1. Developed a dataset that can be used to train a model of Landsat-8 images to Sentinel-2 images.

2. Designed a CNN architecture that is capable to generate Sentinel-2 bands, which have multiple spatial resolutions, using Landsat-8 bands as input.

Materials and methods

Dataset

The Landsat-8 images in this research were acquired from the United States Geological Survey (USGS) EarthExplorer website (United States Geological Survey, 2019). Meanwhile, the Sentinel-2 images were downloaded from the European Space Agency (ESA) Copernicus Open Access Hub website (European Space Agency, 2019). Both Landsat-8 images and Sentinel-2 images that we acquired were Level 2 images. The images were capture by both satellites in the year 2017 and 2018 in the area around San Francisco and near Los Angeles. These locations of the study area were chosen because it has a relatively low cloud cover. These study areas are marked with squares in Figure 5. The tile references for the corresponding study area are given in Table 1. To further ensure the low cloud cover, the
images with more than 10% cloud cover are not included in our dataset.

For each Landsat-8 image, a Sentinel-2 image with the same location and nearest acquisition time was assigned as a target for supervised learning. Because the Landsat-8 images from the given tile reference are bigger than Sentinel-2 images, we cropped the images to match the Sentinel-2 coverage. The bands used for each Landsat-8 images and Sentinel-2 images are the bands that are available from the EarthExplorer and Copernicus Open Access Hub. The bands description is given in tables 2 and 3, respectively, for Sentinel-2 and Landsat-8. After the acquisition and cropping, we converted the pixels value from digital number to Top of Atmosphere (ToA) reflectance value. Because the images are already at Level 2, the conversion is just a scaling by 10^{-4}.

Finally, to form the final dataset, pairs of source and target images were extracted from the processed images. The source images in the final dataset were obtained by tiling the Landsat-8 images to multiple patches of 960 × 960 pixels. Subsequently, the target images were obtained by extracting patches from the Sentinel-2 images with the same geolocation as the Landsat-8 patches. In Figure 6, an example of a data point in the final dataset is illustrated.

| Table 1. Tile reference of the study area. |
|--------------------------------------------|
| Description                  | Landsat-8 (Path/Row) | Sentinel-2 |
| Area around San Francisco    | 44/34                | T11SLU     |
| Area around Los Angeles      | 41/36                | T10SEG     |

Table 2. Sentinel-2 bands.

| Band Name              | Wavelength (μm) | Resolution (m/pixel) |
|------------------------|-----------------|----------------------|
| Band 01 – Coastal Aerosol | 0.43–0.45       | 60                   |
| Band 02 – Blue          | 0.45–0.52       | 10                   |
| Band 03 – Green         | 0.54–0.58       | 10                   |
| Band 04 – Red           | 0.65–0.68       | 10                   |
| Band 05 – Red-edge 1    | 0.70–0.71       | 20                   |
| Band 06 – Red-edge 2    | 0.73–0.75       | 20                   |
| Band 07 – Red-edge      | 0.77–0.79       | 20                   |
| Band 08 – NIR           | 0.79–0.90       | 10                   |
| Band 09 – NIR Narrow    | 0.86–0.88       | 20                   |
| Band 10 – Water Vapour  | 0.94–0.96       | 60                   |
| Band 11 – SWIR 1        | 1.36–1.39       | 60                   |
| Band 12 – SWIR 2        | 1.57–1.66       | 20                   |
| Band 06 – SWIR 2        | 2.10–2.29       | 20                   |

Table 3. Landsat-8 bands.

| Band Name              | Wavelength (μm) | Resolution (m/pixel) |
|------------------------|-----------------|----------------------|
| Band 1 – Coastal Aerosol | 0.43–0.45       | 30                   |
| Band 2 – Blue           | 0.45–0.51       | 30                   |
| Band 3 – Green          | 0.53–0.59       | 30                   |
| Band 4 – Red            | 0.64–0.67       | 30                   |
| Band 5 – NIR            | 0.85–0.88       | 30                   |
| Band 6 – SWIR 1         | 1.57–1.65       | 30                   |
| Band 7 – SWIR 2         | 2.11–2.29       | 30                   |

**Deep learning architecture for Landsat-8 to Sentinel-2 conversion**

Due to the similarity between conversion and image super-resolution problem, we propose a deep-learning-based conversion architecture inspired by the CNN for image super-resolution problem. In image super-resolution research, the architecture used is typically
constructed with a single pathway, which outputs an image with a higher resolution. This architecture cannot be directly applied for Landsat-8 – Sentinel-2 conversion, because the Sentinel-2 image has different resolutions for each band. In particular, the resolution of band 01, 09, and 10 are half of the Landsat-8 resolution, the resolution of band 02, 03, 04, and 08 are three times of the Landsat-8 resolution, and the other bands are 1.5 times of the Landsat-8 resolution. To cope with the variation of the Sentinel-2 image resolution, we modified the typical super-resolution architecture with multiple decoder pathways as depicted in Figure 7. To generate different resolutions, each pathway uses one of the three resolution-adjusting modules: downsampling, upsampling, and 1.5x module.

Figure 6. Dataset illustration.
The downsampling modules in our proposed architecture act as image generators with half resolution of the Landsat-8 image input. We compared two common deep learning modules for downsampling to be used in our proposed architecture: the strided convolution layer and the max pooling layer. The strided convolution layer is similar to the standard convolution layer, except that the kernel moves with a stride of more than one, as illustrated in Figure 8. In this study, we used a stride of two to downsample the input image by half resolution. Different from the strided convolution, max pooling downsamples the input image by taking only the maximum value of each local patch. This operation is illustrated in Figure 9.

Concurrently, the upsampling modules are installed in the proposed architecture to generate an image with
a resolution three times larger than the input. For the upsampling modules, we used two common super-resolution modules: sub-pixel layer (Shi et al., 2016) and transpose convolution layer (Zeiler et al., 2010). Transpose convolution layer was originally named as deconvolution layer by the original authors. However, it is more commonly known as transpose convolution, which is technically more suitable. Figures 10 and 11 respectively illustrate the process of the sub-pixel and transpose convolution layer. The sub-pixel layer works by adding extra dimensions to the convolution kernel (depicted with a red-lined matrix) which generates more feature maps. The upsampled image is formed by aggregating the feature representation as depicted in Figure 10. On the other hand, the transpose convolution layer upsamples the input image by padding the input image with zeroes (depicted as white pixels in Figure 11). The final upsampled image is calculated by performing a convolutional operation to the padded matrix.

The most challenging part in designing the proposed architecture is the resolution-adjusting module for bands with 1.5 times the resolution of the input. A similar challenge has been addressed by Aiazzi et al. (1999) for image fusion. However, their method does not apply to image conversion. Therefore, we designed a module with combination of a downsampling and an upsampling module to generate 1.5 times resolution, named as 1.5x module in Figure 7. Two variants of 1.5x module were considered. The first variant used a sequence of a downsampling sub-module and an upsampling sub-module. The downsampling sub-

---

**Figure 8.** Strided convolution layer 329x220mm (96 x 96 DPI).

**Figure 9.** Max pooling layer 136x64mm (96 x 96 DPI).

**Figure 10.** Sub-pixel 368x118mm (96 x 96 DPI).
module was utilized to generate feature maps with half-resolution. Afterward, an upsampling submodule took the feature maps as input to generate an image with three times resolution of the feature maps. Therefore, the output of the downsampling–upsampling module can generate an image with 1.5 times the resolution of its input. The second variant is the reverse of the first variant, where the first submodule is upsampling, followed by the downsampling sub-module. The downsampling and upsampling sub-module in this module follow the same variation of the downsampling and upsampling module used in other pathways.

In total, we tried eight different architectures in this study. These modules were named: Sub-Stride-DU, Sub-Max-DU, Trans-Stride-DU, Trans-Max-DU, Sub-Stride-UD, Sub-Max-UD, Trans-Stride-UD, and Trans-Max-UD. The first term in the model name refers to the upsampling module used, whether it is sub-pixel (Sub) or transposed convolution (Trans). The second term refers to whether the models used strided convolution (Stride) or max pooling (Max) as the downsampling module. The last term refers to whether the models use a 1.5x module with the downsampling-upsampling (DU) sequence or the upsampling-downsampling (UD) sequence. The architecture of the upsampling, downsampling, and 1.5x module are summarized in Table 4.

**Experiment setting**

All models in this study were trained using the Adam optimization algorithm (Kingma & Ba, 2015) with the default settings. We trained the models with a batch size of 1 for 100 epochs and took the trained models with the best validation loss for the final performance test.

To evaluate the performance of all models, five popular metrics for image quality assessment are used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Peak to Signal Noise Ratio (PSNR), Structural Similarity (SSIM) (Wang et al., 2004), and cross correlation. To assess the similarity between two images, MAE and RMSE are mathematically the most straightforward metrics to assess. MAE and RMSE are calculated as in the equation 1 and 2 respectively, where $x_{ij}$ is the pixel value of the $i^{th}$longitude and $j^{th}$ latitude, $\hat{x}_{ij}$ is the predicted pixel value of the $i^{th}$longitude and $j^{th}$ latitude, and $N$ is the total number of pixels.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |x_{ij} - \hat{x}_{ij}|$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_{ij} - \hat{x}_{ij})^2}$$

In contrast to MAE and RMSE, PSNR is usually more preferred as it takes the peak intensity of the assessed image into account, which makes it more reliable for assessing images with different peak intensity. The formula of PSNR is shown in the equation 3, where $MAX$ is the maximum possible pixel value. In this paper, $MAX$ is set to 10,000, which is a standard denominator for converting Sentinel-2 digital number to reflectance value.

$$PSNR = 20\log_{10} \left( \frac{MAX}{RMSE} \right)$$

SSIM improves the reliability of PSNR further by also considering the structural information of the images. The calculation of SSIM is formulated as in the equation 4, where $\mu_x$ is the average of $x$, $\mu_\hat{x}$ is the average of $\hat{x}$, $\sigma_x$ is the variance of $x$, $\sigma_\hat{x}$ is the variance of $\hat{x}$, $\sigma_{x\hat{x}}$ is the covariance of $x$ and $\hat{x}$, $L$ is the range of pixel values (0 to 10,000), and $k_1$ and $k_2$ are adjustable variables. In this study, $k_1$ and $k_2$ were set to the default values of 0.01 and 0.03, respectively.

$$SSIM = \frac{(2\mu_x\mu_\hat{x} + (k_1L)^2)(2\sigma_{x\hat{x}} + (k_2L)^2)}{\left(\mu_x^2 + \mu_\hat{x}^2 + (k_1L)^2\right)\left(\sigma_x^2 + \sigma_\hat{x}^2 + (k_2L)^2\right)}$$

Additionally, we also used cross correlation as an evaluation metric. Cross correlation measures the
The upsampling, downsampling, and 1.5x module in the CNN architecture for Landsat-8 – Sentinel-2 conversion.

| Upsampling | Sub |
|------------|-----|
| Trans      |     |
| Downsampling |     |
| Stride     |     |
| Max        |     |
| 1.5x DU    |     |
| UD         |     |

geometric distortion between the predicted image and the ground truth image. The formula of cross correlation is given in equation 5.

\[
CC(\hat{x}, x) = \frac{1}{m_i} \sum_{i=1}^{m_1} CCS(\hat{x}_i, x_i)
\]  

(5)

where CCS is the cross correlation for the \( i \)th band in the image, which is calculated as in equation 6.

\[
CCS(A, B) = \frac{\sum_{j=1}^{n} (A_j - \mu_A)(B_j - \mu_B)}{\sqrt{\sum_{j=1}^{n} (A_j - \mu_A)^2 \sum_{j=1}^{n} (B_j - \mu_B)^2}}
\]  

(6)

where \( \mu_A \) is the mean pixel value of band A.

Results and discussion

To compare the performance between all models, we tabulated the MAE, RMSE, PSNR, SSIM, and Cross Correlation of each model in tables 5 and 6. The best performance is marked with red. In general, Trans-Stride-DU achieved the best performance in all metrics, except in cross correlation, which best model is Sub-Stride-DU. The worst overall model was Sub-Max-UD, which is the complete opposite of the best model.

Ablation analysis on the result in tables 5 and 6 revealed that transposed convolution outperformed sub-pixel. Similarly, with ablation analysis, we found that strided convolution is better than max pooling in terms of performance. Surprisingly, the ablation analysis showed that DU models generally have better performance than their UD counterparts. Logically, the information loss in the DU setting should restrict the models to achieve their optimal performance. Our guess is that the DU module behaves similarly to a typical encoder-decoder deep learning architecture, which forces the models to extract only useful information and leads to better performance.

In addition, we inspected the distribution of the predicted image pixel values as well as the performance of all models. In Table 7, we provided the average value of the 5th and 95th percentile of the ground truth and the predicted values from all models. In general, the predicted values by models with a downsampling-upsampling module have a similar range to the ground truth pixels value, except for band 10. In contrast, the difference in range was noticeable for the predicted values from models with an upsampling-downsampling module. The minimum and maximum predicted values by all models were not reported because they were suspected to be predictions of noisy pixels.
In Table 8, we summarized the mean and standard deviation of the ground truth and the predicted values from all models, excluding data below 5th percentile and above 95th percentile. Generally, all models have noticeably different statistics than the ground truth for band 10. This suggests that all models were not able to predict band 10 well. Although the MAE and RMSE for the band 10 predictions in Table 5 seemed to be smaller than the other bands, it should be noticed that the range of pixel values for band 10 are much smaller than other bands, as displayed in Table 7, leading to the smaller MAE and RMSE. The failure of band 10 prediction was most likely caused by the fact that band 10 mainly captures the wavelength range of the cirrus cloud. Due to the absence of a band with a similar wavelength range in Landsat-8 data we obtained, it should be expected that no model can predict band 10 well.

Conversely, in other bands, the mean of the predicted values by all models was similar to the ground truth. However, the standard deviation of the UD models deviated significantly from the standard deviation of the ground truths. In contrast, the standard deviation value of the DU models was similar to the standard deviation of the ground truths. Therefore, the distribution of the predicted pixel values by down-sampling-upsampling models can be concluded as similar to the ground truth pixel values, indicating a successful prediction. However, it was not the case...
for the predicted pixel values by the UD models, which statistically seemed to have a different distribution than the ground truths.

To confirm the displayed trend in tables 7 and 8, we plotted the predicted pixel values distributions of the best and the worst model in a histogram, which are displayed in Appendix A (Figure A.1 to A.13). The histogram plots showed that the distributions of predicted pixel values by the best model generally have a similar shape, but smoother than the ground truth distributions. The only exception was the predicted distribution of band 10, which indicated a failure in predicting band 10. In contrast, the distributions of predicted pixel values by the worst model were noticeably different from the ground truth distribution of all bands. This confirms the displayed trend in tables 7 and 8.

Additionally, we confirmed the prediction of the best model (Trans-Stride-UD) with visual comparison to the respective ground truth. For this comparison, we extracted patches of 64 x 64 with the best MAE, the median MAE, and the worst MAE, as displayed in tables 9 and 10. From the visual inspection, it can be concluded that Trans-Stride-UD can generally capture the structure of the ground truth images. The prediction of band 02, 03, 04, and 08 was blurrier than the other bands, but still reflected the general shape of the ground truth images. By comparing the best, median, and worst prediction, we could deduce that the best conversion model tended to predict images with less texture.
Table 9. Visual comparison of the best model (Trans-Stride-DU) prediction to the corresponding ground truth for Band 01, 02, 03, 04, 05, 06, and 07.

| Band | 01 | 02 | 03 | 04 | 05 | 06 | 07 |
|------|----|----|----|----|----|----|----|
| Best MAE | Ground Truth | ![Image](image1) | ![Image](image2) | ![Image](image3) | ![Image](image4) | ![Image](image5) | ![Image](image6) | ![Image](image7) |
| Prediction | ![Image](image8) | ![Image](image9) | ![Image](image10) | ![Image](image11) | ![Image](image12) | ![Image](image13) | ![Image](image14) |
| Median MAE | Ground Truth | ![Image](image15) | ![Image](image16) | ![Image](image17) | ![Image](image18) | ![Image](image19) | ![Image](image20) | ![Image](image21) |
| Prediction | ![Image](image22) | ![Image](image23) | ![Image](image24) | ![Image](image25) | ![Image](image26) | ![Image](image27) | ![Image](image28) |
| Worst MAE | Ground Truth | ![Image](image29) | ![Image](image30) | ![Image](image31) | ![Image](image32) | ![Image](image33) | ![Image](image34) | ![Image](image35) |
| Prediction | ![Image](image36) | ![Image](image37) | ![Image](image38) | ![Image](image39) | ![Image](image40) | ![Image](image41) | ![Image](image42) |

Table 10. Visual comparison of the best model (Trans-Stride-DU) prediction to the corresponding ground truth for Band 08, 8A, 09, 10, 11, and 12.

| Band | 08 | 8A | 09 | 10 | 11 | 12 |
|------|----|----|----|----|----|----|
| Best MAE | Ground Truth | ![Image](image43) | ![Image](image44) | ![Image](image45) | ![Image](image46) | ![Image](image47) | ![Image](image48) |
| Prediction | ![Image](image49) | ![Image](image50) | ![Image](image51) | ![Image](image52) | ![Image](image53) | ![Image](image54) |
| Median MAE | Ground Truth | ![Image](image55) | ![Image](image56) | ![Image](image57) | ![Image](image58) | ![Image](image59) | ![Image](image60) |
| Prediction | ![Image](image61) | ![Image](image62) | ![Image](image63) | ![Image](image64) | ![Image](image65) | ![Image](image66) |
| Worst MAE | Ground Truth | ![Image](image67) | ![Image](image68) | ![Image](image69) | ![Image](image70) | ![Image](image71) | ![Image](image72) |
| Prediction | ![Image](image73) | ![Image](image74) | ![Image](image75) | ![Image](image76) | ![Image](image77) | ![Image](image78) |
**Conclusion**

This study shows a promising possibility of developing a machine learning model for converting images of Landsat-8 to Sentinel-2. Our best model, Trans-Stride-DU, successfully predicted images that are quantitatively and qualitatively similar to the actual Sentinel-2 images. This model used a transposed convolution as the upsampling module, a strided convolution as the downsampling module, and a sequence of downsampling and upsampling module to generate Sentinel-2 bands with 1.5 times the resolution of Landsat-8 bands.

The models introduced in this study can conceptually be applied to convert images from and to any satellite. Therefore, the studies following this research can explore the use of images from different satellites for conversion. Another interesting research direction is to utilize unsupervised learning, which can alleviate the rigorous data preparation that is necessary for the supervised learning model.

**Acknowledgments**

This study is funded by Directorate of Research and Community Service, Directorate General of Research and Development, Indonesian Ministry of Research, Technology and Higher Education as a part of 2019 Penelitian Terapan Unggulan Perguruan Tinggi Research Grant. Part of the experiments was run using NVIDIA Tesla P100 supported by NVIDIA - BINUS AI R&D Center.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Funding**

This work was supported by the Directorate of Research and Community Service, Directorate General of Research and Development, Indonesian Ministry of Research, Technology and Higher Education [225/SP2H/LT/DRPM/2019].

**ORCID**

Sani M. Isa http://orcid.org/0000-0003-4102-7368
Tjeng Wawan Cenggoro http://orcid.org/0000-0002-9872-9646

**Data availability statement**

The code and data that support the findings of this study are openly available at https://github.com/wawancenggoro/cnn_18_s2.

**References**

Aiazzì, B., Alparone, L., Barducci, A., Baronti, S., & Pippi, I. (1999). Multispectral fusion of multisensor image data by the generalized Laplacian pyramid. *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS’99* (Cat. No.99CH36293) 2, 1183–1185. Hamburg, Germany. https://doi.org/10.1109/IGARSS.1999.774572

Belgiu, M., & Stein, A. (2019). Spatiotemporal image fusion in remote sensing. *Remote Sensing, 11*(7), 818. https://doi.org/10.3390/rs11070818

Cao, Y., Niu, X., & Dou, Y. (2016). Region-based convolutional neural networks for object detection in very high resolution remote sensing images. *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD),* 548–554. Changsha, China.

Chabib, S., Liu, H., Gu, Y., & Yao, H. (2017). Deep feature fusion for VHR remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing, 55*(8), 4775–4784. https://doi.org/10.1109/TGRS.2017.2700322

Cheng, G., Ma, C., Zhou, P., Yao, X., & Han, J. (2016). Scene classification of high resolution remote sensing images using convolutional neural networks. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS),* 767–770. Beijing, China.

Cheng, G., Yang, C., Yao, X., Guo, L., & Han, J. (2018). When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Transactions on Geoscience and Remote Sensing, 56*(5), 2811–2821. https://doi.org/10.1109/TGRS.2017.2783902

Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J. C., Skakun, S. V., & Justice, C. (2018). The harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment, 219*(August), 145–161. https://doi.org/10.1016/j.rse.2018.09.002

Collins, C. B., Beck, J. M., Bridges, S. M., Rushing, J. A., & Graves, S. J. (2017). Deep learning for multisensor image resolution enhancement. *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery,* 37–44. Los Angeles, CA, USA. https://doi.org/10.1145/3149808.3149815

Dian, R., Li, S., Guo, A., & Fang, L. (2018). Deep hyperspectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems,* 29(11), 5345–5355. https://doi.org/10.1109/TNNLS.2018.2798162

Dong, C., Loy, C. C., & He, K. (2014). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(2), 1–14. https://doi.org/10.1109/TPAMI.2015.2439281

European Space Agency. (2019). *Copernicus open access hub.* https://scihub.copernicus.eu/

Gao, F., Masek, J., Schwaller, M., & Hall, F. (2006). On the blending of the landsat and MODIS surface reflectance: Predicting daily landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing, 44*(8), 2207–2218. https://doi.org/10.1109/TGRS.2006.872081

Ghamisi, P., Gloaguen, R., Atkinson, P. M., Benediktsson, J. A., Rasti, B., Yokoya, N., Wang, Q., Hölle, B., Bruzzone, L., Bovolo, F., Chi, M., & Anders, K. (2019). Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine, 7*(1), 6–39. https://doi.org/10.1109/MGRS.2018.2890023

Gu, Y., Wang, Y., & Li, Y. (2019). A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection. *Applied Sciences,* 9(10), 2110. https://doi.org/10.3390/app9102110
Han, L., Li, P., Bai, X., Grecos, C., Zhang, X., & Ren, P. (2020). Cohesion intensive deep hashing for remote sensing image retrieval. Remote Sensing, 12(1), 1–12. https://doi.org/10.3390/RS12010101

Hilker, T., Wulder, M. A., Coops, N. C., Linke, J., McDermid, G., Masek, J. G., Gao, F., & White, J. C. (2009). A new data fusion model for high-spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS. Remote Sensing of Environment, 113(8), 1613–1627. https://doi.org/10.1016/j.rse.2009.03.007

Huang, W., Xiao, L., Wei, Z., Liu, H., & Tang, S. (2015). A new pan-sharpening method with deep neural networks. IEEE Geoscience and Remote Sensing Letters, 12(5), 1037–1041. https://doi.org/10.1109/LGRS.2014.2376034

Jia, Y., Ge, Y., Chen, Y., Li, S., Heuvelink, G., & Ling, F. (2019). Super-resolution land cover mapping based on the convolutional neural network. Remote Sensing, 11(15), 1815. https://doi.org/10.3390/rs11151815

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. The International Conference on Learning Representations 2015. San Diego, CA, USA. http://arxiv.org/abs/1412.6980

Lanaras, C., Boucas-Dias, J., Galliani, S., Baltsevias, E., & Schindler, K. (2018). Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. ISPRS Journal of Photogrammetry and Remote Sensing, 146, 305–319. https://doi.org/10.1016/j.isprsjprs.2018.09.018

Li, P., Han, L., Tao, X., Zhang, X., Grecos, C., Plaza, A., & Ren, P. (2020). Hashing nets for hashing: A quantized deep learning to hash framework for remote sensing image retrieval. IEEE Transactions on Geoscience and Remote Sensing, 58(10), 1–15. https://doi.org/10.1109/TGRS.2020.2981997

Li, P., Ren, P., Zhang, X., Wang, Q., Zhu, X., & Wang, L. (2018). Region-wise deep feature representation for remote sensing images. Remote Sensing, 10(6), 1–14. https://doi.org/10.3390/rs10060871

Liu, Q., Hang, R., Song, H., & Li, Z. (2018). Learning Multiscale Deep Features for High-Resolution Satellite Image Scene Classification. IEEE Transactions on Geoscience and Remote Sensing, 56(1), 117–126. https://doi.org/10.1109/TGRS.2017.2743243

Liu, W., & Lee, J. (2019). An efficient residual learning neural network for hyperspectral image superresolution. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(4), 1240–1253. https://doi.org/10.1109/JSTARS.2019.2901752

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440. Boston, MA, USA.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. ISPRS Journal of Photogrammetry and Remote Sensing, 152(November2018), 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

Masi, G., Cozzolino, D., Verdoliva, L., & Scarpa, G. (2016). Pan-sharpening by convolutional neural networks. Remote Sensing, 8(7), 594. https://doi.org/10.3390/rs8070594

Masi, G., Cozzolino, D., Verdoliva, L., & Scarpa, G. (2017). CNN-based pansharpening of multi-resolution remote-sensing images. 2017 Joint Urban Remote Sensing Event (JURSE), 1–4. Dubai, United Arab Emirates. https://doi.org/10.1109/JURSE.2017.7924534

Palsson, F., Sveinsson, J. R., & Ulfsaroos, M. O. (2017). Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network. IEEE Geoscience and Remote Sensing Letters, 14(5), 639–643. https://doi.org/10.1109/LGRS.2017.2668299

Pouliot, D., Latifovic, R., Pasner, J., & Dufè, I. (2018). Landsat super-resolution enhancement using convolution neural networks and Sentinel-2 for training. Remote Sensing, 10(3), 394. https://doi.org/10.3390/rs10030394

Ran, Q., Xu, X., Zhao, S., Li, W., & Du, Q. (2019). Remote sensing images super-resolution with deep convolution networks. Multimedia Tools and Applications, 79(13), 1–17. https://doi.org/10.1007/s11042-018-7091-1

Ren, Y., Zhu, C., & Xiao, S. (2018). Small object detection in optical remote sensing images via modified faster R-CNN. Applied Sciences, 8(5), 813. https://doi.org/10.3390/app8050813

Scarpa, G., Gargiulo, M., Mazza, A., & Gaetano, R. (2018). A CNN-based fusion method for feature extraction from sentinel data. Remote Sensing, 10(2), 1–20. https://doi.org/10.3390/rs10020236

Shao, Z., & Cai, J. (2018). Remote sensing image fusion with deep convolutional neural network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(5), 1656–1669. https://doi.org/10.1109/JSTARS.2018.2805923

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1874–1883. Las Vegas, NV, USA. https://doi.org/10.1109/CVPR.2016.207

United States Geological Survey. (2019). EarthExplorer. https://earthexplorer.usgs.gov/

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P., & others. (2004). Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Wei, Y., Yuan, Q., Shen, H., & Zhang, L. (2017). Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. IEEE Geoscience and Remote Sensing Letters, 14(10), 1795–1799. https://doi.org/10.1109/LGRS.2017.2736020

Xing, Y., Wang, M., Yang, S., & Jiao, L. (2018). Pansharpening via deep metric learning. ISPRS Journal of Photogrammetry and Remote Sensing, 145(A), 165–183. https://doi.org/10.1016/j.isprsjprs.2018.01.016

Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., & Paisley, J. (2017). PanNet: A deep network architecture for pan-sharpening. Proceedings of the IEEE International Conference on Computer Vision, 1753–1761. Venice, Italy. https://doi.org/10.1109/ICCV.2017.193

Yang, J., Zhao, Y. Q., & Chan, J. C. W. (2018). Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network. Remote Sensing, 10(5), 800. https://doi.org/10.3390/rs10050800

Yuan, Q., Wei, Y., Meng, X., Shen, H., & Zhang, L. (2018). A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(3), 978–989. https://doi.org/10.1109/JSTARS.2018.2794888

Zeller, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010). Deconvolutional networks. Computer Vision and
Appendix A. Pixel Value Distribution Plotted in Histogram

Figure A.1. Pixel value distribution of band 1 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.2. Pixel value distribution of band 2 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.3. Pixel value distribution of band 3 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.4. Pixel value distribution of band 4 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.5. Pixel value distribution of band 5 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.6. Pixel value distribution of band 6 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.7. Pixel value distribution of band 7 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.8. Pixel value distribution of band 8 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.9. Pixel value distribution of band 8A plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.10. Pixel value distribution of band 9 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.11. Pixel value distribution of band 10 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.12. Pixel value distribution of band 11 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.
Figure A.13. Pixel value distribution of band 12 plotted in histogram: (a) ground truth; (b) prediction by using Trans-Stride-DU; (c) prediction by using Sub-Max-UD.