Machine Learning with Applications 9 (2022) 100381

Contents lists available at ScienceDirect

Machine Learning with Applications

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

Hybrid U-Net: Semantic segmentation of high-resolution satellite images to detect war destruction

Shima Nabiee [a,*], Matthew Harding [b], Jonathan Hersh [c], Nader Bagherzadeh [a]

[a] Department of Electrical Engineering and Computer Science, Henry Samueli School of Engineering, University of California in Irvine, USA
[b] Economics Department, University of California in Irvine, USA
[c] George L. Argyros School of Business and Economics, Chapman University, USA

ARTICLE INFO

Keywords:
War destruction detection
Semantic segmentation
U-Net
High-resolution satellite images

ABSTRACT

Destruction caused by violent conflicts play a big role in understanding the dynamics and consequences of conflicts, which is now the focus of a large body of ongoing literature in economics and political science. However, existing data on conflict largely come from news or eyewitness reports, which makes it incomplete, potentially unreliable, and biased for ongoing conflicts. Using satellite images and deep learning techniques, we can automatically extract objective information on violent events. To automate this process, we created a dataset of high-resolution satellite images of Syria and manually annotated the destroyed areas pixel-wise. Then, we used this dataset to train and test semantic segmentation networks to detect building damage of various size. We specifically utilized a U-Net model for this task due to its promising performance on small and imbalanced datasets. However, the raw U-Net architecture does not fully exploit multi-scale feature maps, which are among the important factors for generating fine-grained segmentation maps, especially for high-resolution images. To address this deficiency, we propose a multi-scale feature fusion approach and design a multi-scale skip-connected Hybrid U-Net for segmenting high-resolution satellite images. In our experiments, U-Net and its variants demonstrated promising segmentation results to detect various war-related building destruction. In addition, Hybrid U-Net resulted in significant improvement in segmentation performance compared to U-Net and other baselines. In particular, the mean intersection over union and mean dice score improved by 7.05% and 8.09%, respectively, compared to those in the raw U-Net.

1. Introduction

The empirical analysis has been attracting attention in recent years in efforts to understand the profile, causes, actors, and dynamics of conflicts. Locating and documenting building destruction caused by armed conflict are core elements of conflict analysis which plays a significant role in analyzing the dynamics of wars (Fisher et al., 2000; Jabri, 1996; Mason & Rychard, 2005). Geo-spatial data have made it possible to create panel data on violent events, which are often based on news and human rights sources. This has led researchers to develop theories to help understand whether, why, and how a conflict is escalating, intensifying, decreasing, spreading, contracting, or in stalemate (UK Government's Stabilisation Unit, 2017).

These reported war events data have some shortcomings. First, most practical applications rely on labor-intensive image analysis and require considerable time between observation and processing (Knoth, Slimani, Appel, & Pebesma, 2018). This manual procedure can be time and labor intensive, and may lead to delayed response and further catastrophe. Second, these data may be biased, due to the absence of data about inaccessible areas (Witmer, 2015; Wolfinbarger & Wyndham, 2011). For these concerns, an automatic remote sensing framework to monitor and detect violent events is essential (Avtar et al., 2021; Braun, 2019; Kishi, 2021; Quinn et al., 2018). Many authors have employed automatic image processing techniques on remote sensing data intending to find and analyze the impacts of combat, such as damaged and destroyed building structures (Marx & Loboda, 2013; Pagot & Pesaresi, 2008; Sulik & Edwards, 2010).

In particular, authors in Mueller, Groeger, Hersh, Matranga, and Serrat (2021) have addressed the problem of identifying building damage caused by the Syrian civil war using pre-and post-destruction satellite images. They developed a patch-wise change detection framework to classify patches as destroyed or intact. Despite their success in patch-wise binary classification, the intensity and size of the damage using this approach could not be inferred. The authors have focused on the identification of building damage in urban areas, even though detecting damages to all types of structures, such as roads and farmlands, is necessary for the accurate assessment of conflict dynamics. Additionally, acquiring pre-destruction images is costly. To address these problems, we propose to apply semantic segmentation techniques

* Corresponding author.
E-mail addresses: snabiee@uci.edu (S. Nabiee), harding1@uci.edu (M. Harding), hersh@chapman.edu (J. Hersh), nader@uci.edu (N. Bagherzadeh).

https://doi.org/10.1016/j.mlwa.2022.100381
Received 2 December 2021; Received in revised form 8 June 2022; Accepted 6 July 2022
Available online 9 July 2022

to spot damages using only post-destruction satellite images. As we will discuss later, U-Net (Ronneberger, Fischer, & Brox, 2015), an encoder–decoder network for semantic segmentation of medical images, has been very successful in the segmentation of images containing various size objects on complex backgrounds even with a limited number of training images.

In this paper, we show the effectiveness of U-Net for detecting war destruction. We further enhance the current U-Net model to perform better on segmentation of high-resolution satellite images, proposing a novel multi-scale feature fusion schema as an extension to U-Net while preserving its symmetry. Additionally, we collected and pixel-wise annotated a dataset of high-resolution satellite images of Syria and tested various models on it. According to our experiments, not only U-Net is very effective in segmenting satellite images of destruction, but also our proposed architecture yields significant performance gain over U-Net and its other variations.

The main contributions of this paper are as follows:

1. Introducing a dataset of high-resolution satellite images capturing destroyed areas of Syria. Each image comes with its corresponding segmentation map which we manually created.
2. Proposing Hybrid U-Net, a symmetric multi-scale feature fusion schema inspired by FCN to use deep coarse feature maps, resulting in the prediction of more refined segmentation maps.
3. Performing comprehensive experiments, demonstrating the great capability of U-Net and its variations in detecting war destruction.

2. Technical background

Semantic segmentation methods in deep learning have emerged in recent years with two representative network architectures: FCN-based and encoder–decoder architectures. Fully convolutional networks (FCN) (Long, Shelhamer, & Darrell, 2015) replace the fully connected layers with convolutional layers and extend the network by adding learnable upsampling layers. It deploys skip connections and deconvolution layers to combine coarse information from deep layers with fine information from shallow layers for an end-to-end pixel-wise classification. Although FCN omits fully connected layers thus preserving the spatial information crucial for localization, the results of its upsampling are relatively fuzzy and insensitive to the details of the image, resulting in the segmentation results not being fine enough.

Encoder–decoder frameworks such as U-Net (Ronneberger et al., 2015) and SegNet (Badrinarayanan, Kendall, & Cipolla, 2017) are also widely used for semantic segmentation. Like autoencoders, they encode a hierarchical representation of an image in compressed latent space and decode that representation into a segmentation map with the same size as the input image. U-Net utilizes symmetric skip connections to connect feature maps of the encoder sub-network to the upsampled feature maps of the decoder sub-network. Due to its symmetry, it has a satisfactory performance on small datasets. More importantly, combining multi-scale feature maps has proved to be effective in recognizing and localizing objects at different scales, even for images with complex backgrounds (Lin et al., 2017). Nevertheless, U-Net has failed to fully utilize information from different scales. To address this issue, many modifications to U-Net have been proposed.

For medical image segmentation, UNet++ (Zhou, Siddiquee, Tajbakhsh, & Liang, 2018) has introduced nested and dense skip connections to raw U-Net. However, it still does not explore sufficient information from multi-scale feature maps. Taking advantage of full-scale skip connections, UNet 3+ (Huang et al., 2020), overcame this limitation at the cost of considerable computational complexity. Cao et al. (2021) proposed Swin-Unet by replacing the CNN blocks in U-Net with Swin transformer blocks (Liu et al., 2021) to take advantage of the feature representation learning power of vision transformers. U-Net variants have also proved to be successful in semantic segmentation of high-resolution satellite images. Authors in Korznikov et al. (2021) applied a U-shaped CNN architecture to very high-resolution RGB satellite images for tree recognition. MACU-Net (Li, Duan, Zheng, Zhang, & Atkinson, 2021) utilized attention blocks to combine previous encoder and decoder multi-scale feature maps and supplied each decoder stage with those rich features.

In fact, many studies have proposed to improve U-Net by minimizing the semantic gap between the encoder and decoder sub-networks using different skip connection schemes and fusion techniques, among which are residual skip connections and dense skip connections (Siddique, Paheding, Elkin, & Devabhaktuni, 2021). Residual skip connections (He, Zhang, Ren, & Sun, 2016) fuse the skipped input feature map of each ResNet block with the output of its last layer via element-wise addition. Each block can have one or multiple CNN layers. In Zhang, Liu, and Wang (2018) authors used residual blocks for each encoder and decoder stage with double-layer skips. Zhang et al. (2019) proposed residual skip connection between feature maps of the same level of the encoder and decoder sub-networks. The main motivation to use ResNet blocks is mostly to overcome the difficulty in training deeper neural networks by helping the vanishing gradient problem.

Many also employed dense skip connections along with DenseNet (Huang, Liu, Van Der Maaten, & Weinberger, 2017) blocks in order to fully exploit the previous layers' information. While the ResNet model allows for deeper networks, it does not preserve feature maps from multiple prior layers. In dense block implementation, every layer in a block receives the feature maps from all of its preceding layers, which are then combined via channel-wise concatenation. A multi-scale U-Net using dense skip connections is studied in Zhang, Jin, Xu, Xu, and Zhang (2018). They use dense blocks in every stage of encoder and decoder and propose three different multi-scale dense connections for encoder, decoder, and across them. A symmetric lightweight multi-scale U-Net (Tarasiewicz, Nalepa, & Kawulok, 2020) proposed to increase the feature propagation and captured spatial context by using dense blocks along with inception modules. They outperformed vanilla U-Net with better generalization over unseen data. U-Net++ is another example of a densely skip-connected design resulting in a nested U-Net architecture. Dense skip connections have proved to improve the raw U-Net's performance and also allow for deeper U-Net models (Siddique et al., 2021).

Other fusion techniques and skip-connection designs have also been proposed. In Phan, Kim, Yang, Lee, et al. (2021) authors proposed a multitask U-Net-based architecture by redesigning the decoder for each task. The semantic segmentation decoder uses a similar skip connection to those of our version 1 skip connection schema. Authors in Lee et al. (2022) designed two non-symmetrical U-Net variants. In their initial design, every decoder layer receives skip connections from all encoder layers, leading to high computational cost, resolved by using a pruned version of it. In Yuan, Liu, and Wang (2019) authors used attention unit alongside multi-scale feature maps from all layers of the encoder and concatenated it with the deepest layer semantic information.

Despite many existing multi-scale skip connected U-Net variations, to the best of our knowledge there is no similar skip connected structure to the one we propose. We mainly aimed to keep the U-Net's symmetry to be able to fully benefit from its power and be able to recover deeper layers' information by designing a skip-connected sub-network. We do not extend the connections between the main encoder and decoder sub-networks to supply more information to the decoder. We argue that although feeding each decoder stage with multiple feature maps from previous stages leads to better performance compared to raw U-Net, very useful deep coarse information is still lost. As we will see, this leads to inaccuracy in recognizing small objects.

3. Dataset

To segment Syria's satellite images searching for various size war-inflicted destruction, we created satellite images of destroyed areas from four cities of Syria. These cities have been under heavy weaponry

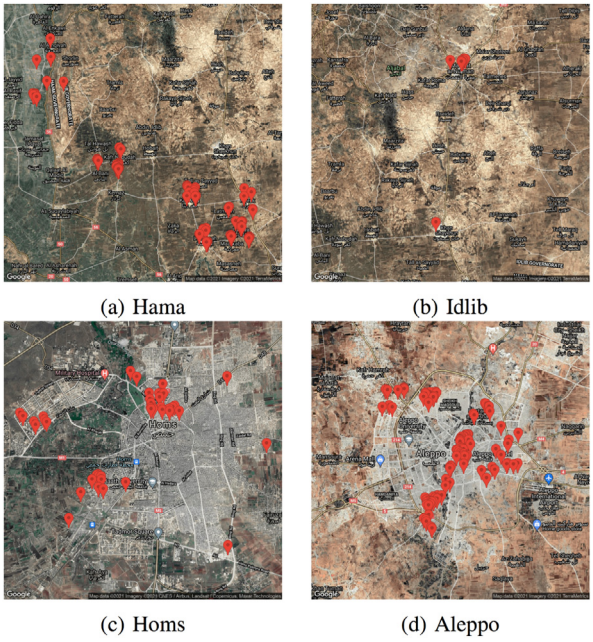

Fig. 1. Distribution of the satellite images' geo-locations for each city.[1]

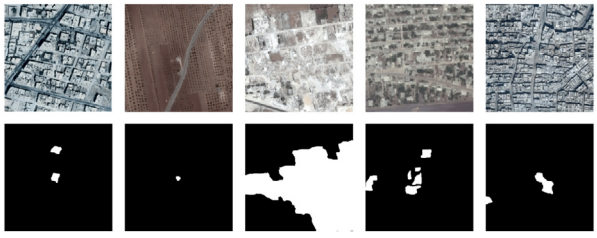

Fig. 2. Satellite images and ground truth masks. Destroyed ground truth areas are shown in white.

attack during the Syrian civil war. We identified areas containing at least one destroyed spot. Fig. 1 shows the geolocations of the 252 selected spots inside each city.

To obtain the post-event satellite images using these coordinates we used the Google Maps Static API. This service creates a map based on URL parameters sent through a standard HTTP request and returns the map as an RGB image. For each HTTP request, a specific location on the map, size of the image, zoom level, type of the map, and placement of optional markers at locations on the map should be defined. Setting the image size to 640 and the scale to 2 returns a 1280×1280 image that is centered at the requested coordinates. We set the zoom level and type of the map to 17 and satellite, respectively. Images are then cropped to eliminate the Google mark, decreasing their size to 1248×1248 pixels.

Finally, the ground truth labels are created to train and test semantic segmentation networks. Each image has been pixel-wise-annotated manually by creating black and white mask images. The fraction of destruction for each image is calculated by dividing the size of the positive class, i.e. number of the white pixels, by the size of the image. Table 1 shows a summary of the number of images per district and the average ratio of the destroyed areas per image. The final dataset contains 252 RGB images and their corresponding BW labels with size 1248×1248 pixels. Several examples of the images and their ground truth masks are demonstrated in Fig. 2. Dataset and codes can be found at Nabiee, Hersh, Harding, and Bagherzadeh (2021).

[1] Google Maps Geocoding API.

4. Damage recognition network

Fig. 3(a) illustrates the structure of U-Net, the backbone of our network architecture. It consists of an encoder sub-network followed by a decoder sub-network. The decoder sub-network up-samples features using a transposed convolution that corresponds to each down-sampling stage in the encoder sub-network. The upsampled features are fused with the corresponding feature maps from the encoder that have the same resolution.

In further detail, each stage of the encoder sub-network consists of a sequence of two 3×3 convolution layers with Rectified Linear Unit (ReLU) activation functions, followed by a max-pooling operation with a pooling size of 2×2. After each down-sampling stage, the number of filters in the convolutional layers is doubled. The filter size of each layer is indicated below each convolution block in Fig. 3(a). This sequence is repeated four times, creating four semantically-rich feature maps in multiple scales. Each stage of the decoder sub-network first up-samples the feature map using a 2×2 transposed convolution operation, halving the number of filters in each stage. Then again a sequence of two 3×3 convolution+ReLU operations is performed. After repeating this operation four times, a 1×1 convolution operation followed by a Softmax layer is performed to generate the final segmentation mask.

As mentioned earlier, the performance of U-Net is limited by its scarce utilization of information streams. Being excessively restrictive, U-Net fuses only the same-scale feature maps from the encoder and decoder sub-networks, which causes a large semantic gap between these feature maps. On the other hand, U-Net does not fully utilize the feature information of the deep and coarse layers, which are highly important for extracting the global structure. To address these shortcomings, we propose a multiscale skip-connected architecture that fuses coarse semantic feature maps and fine appearance feature maps from both the decoder and encoder sub-networks. Exploiting deep layers feature maps efficiently is not only substantial for better localization, but also for detection of the smaller size destroyed spots.

Likewise, authors in Xie and Tu (2015) proposed a holistically nested edge detection network that uses multi-scale and multi-level features to produce object boundary segmentation maps. They adopted a pre-trained VGGNet as the encoder of an FCN-based architecture and fused upsampled multi-scale features from all encoder stages at the output node. Adding links from deep layers to the output led to a significant gain in performance over the traditional FCNs. We use a similar idea to design our skip-network architecture, or the new decoder sub-network, except we use the U-Net multi-scale feature maps as the encoded features. This is due to the facts that (a) U-Net is proven to perform better than the FCN (Ahmed, Ahmad, Khan, & Asif, 2020), especially in the absence of pre-trained models, and (b) symmetric encoder–decoder skip-connected networks converge faster than the non-symmetric ones and achieve better results, even with small dataset (Mao, Shen, & Yang, 2016). Accordingly, we design the additional decoder sub-network such that the U-Net's symmetry is preserved, proposing Hybrid U-Net. The extended decoder sub-network fuses the same-level feature maps from both the encoder and decoder sub-networks and creates full-resolution feature maps from different levels. We then fuse these decoder sub-networks' full-resolution feature maps in a concatenation envelope to create a final semantic segmentation map. Our proposed Hybrid U-Net architecture is demonstrated in Fig. 3(b).

Similar to raw U-Net, each encoder stage consists of two consecutive CNN+ReLU layers. The pooling layer in each X_{En} level halves the feature stacks' height and width. This dimension reduction is compensated for by doubling the number of channels. Then, the encoded coarse feature maps from X^5_{En} are fed to the decoder sub-network. Each decoder layer, X^n_{De}, fuses the upsampled features from the previous stage with the corresponding encoder layer (X^n_{En}), followed by two consecutive CNN+ReLU layers. Using extra skip connections, these

Table 1
Dataset geo-locations and information.

Governorate	Approx. covered area (km^2)	Sub-district	Lat. range	Long. range	No. of images	Fraction of destructions (%)
Hama	36.5	Muhradah	35.29–35.38	36.60–36.70	51	3.27
		As Suqaylabiyah Al	35.40–35.57	36.36–36.49	36	3.92
		Haffah	35.49–35.50	36.35–36.36	7	0.89
Idlib	3.5	Maarrat al-Numan	35.65–35.65	36.66–36.69	7	1.49
		Khan Shaykhun	35.43–35.44	36.64–36.65	2	0.86
Homs	17.5	Homs	34.68–34.75	36.66–36.75	45	9.32
Aleppo	40.5	Mount Simeon	36.15–36.24	37.09–37.20	104	16.71

Table 2
Convolution layers details.

Layer	Size	Layer	Size
$X^1_{En(De)}$	1248 × 1248 × 16	X^1_{Skip}	624 × 624 × 48
$X^2_{En(De)}$	624 × 624 × 32	X^2_{Skip}	312 × 312 × 96
$X^3_{En(De)}$	312 × 312 × 64	X^3_{Skip}	156 × 156 × 192
$X^4_{En(De)}$	156 × 156 × 128	X^4_{Skip}	78 × 78 × 384
X^5_{En}	78 × 78 × 256	Final	1248 × 1248 × 128

feature maps are also utilized in the supplementary sub-network. Each layer of the proposed sub-network first fuses the input to each down-sampling layer with the output of the corresponding up-sampling layer. To achieve better segmentation results, we used concatenation as the fusion method to allow the network to learn the weighed fusion of the features. We up-sampled the resolution of these extended feature maps, X^n_{Skip}, to be the same as the resolution of the input image using transposed convolutions.

We formulated the skip pathway to the output concatenation envelope as follows: let X^n_{En} denotes the output of the encoder where n indexes the down-sampling layer alongside the encoder, and let X'^n_{En} and X^n_{De} be its corresponding pooled encoded features and decoder layer, respectively. The stack of feature maps represented by X^n is computed as:

$$X^n = \begin{cases} \mathcal{T}\ ([X'^n_{En}, X^{n+1}_{De}], stride = 2^n), & n < 4 \\ \mathcal{T}\ ([X'^n_{En}, X^{n+1}_{En}], stride = 2^n), & n = 4 \end{cases} \quad (1)$$

Where $\mathcal{T}$ denotes the transposed convolution with the given stride.

With these added skip connections, the final layer is provided with U-Net's last-stage feature map (X^1_{De}) and with the upsampled intermediate feature maps, X^ns, which are combined encoder and decoder sub-networks' features (see Table 2). Once the envelope is concatenated with all the full-resolution feature maps, the predictions are performed by jointly considering the feature maps from the five streams. That is, the concatenation envelope is followed by a CNN with 128 channels and a sum layer. The final scores are obtained after applying the Softmax layer. Finally, the binary cross-entropy loss function is used for end-to-end training. For each image, the binary cross-entropy can be expressed as:

$$BCE\ Loss = -\sum_{j \in Y_+} \log P(y_j = 1|X) - \sum_{j \in Y_-} \log P(y_j = 0|X) \quad (2)$$

Where $Pr(y_j = 1|X) \in [0, 1]$ is computed using Softmax function on the activation value at pixel j.

5. Experiment details

5.1. Models

To evaluate the effectiveness of our proposed network, we used the U-Net, U-Net++ (Zhou et al., 2018), and MACU-Net (Li et al., 2021) models as our baseline for U-net and its multi-scale variants to show the capability of the added skip sub-network. As mentioned, U-Net++ aims to modify U-Net by bridging the semantic gap between the feature maps of the encoder and the decoder prior to their fusion. To do so, they used a series of nested, dense skip connections to connect the encoder and decoder sub-networks. MACU-Net is an adaptation of the raw U-Net for semantic segmentation of high-resolution satellite images. It utilizes asymmetric convolution block (ACB) along with multiscale skip connections and attention blocks to enhance the representation power of convolution layers and combine semantic features of different levels.

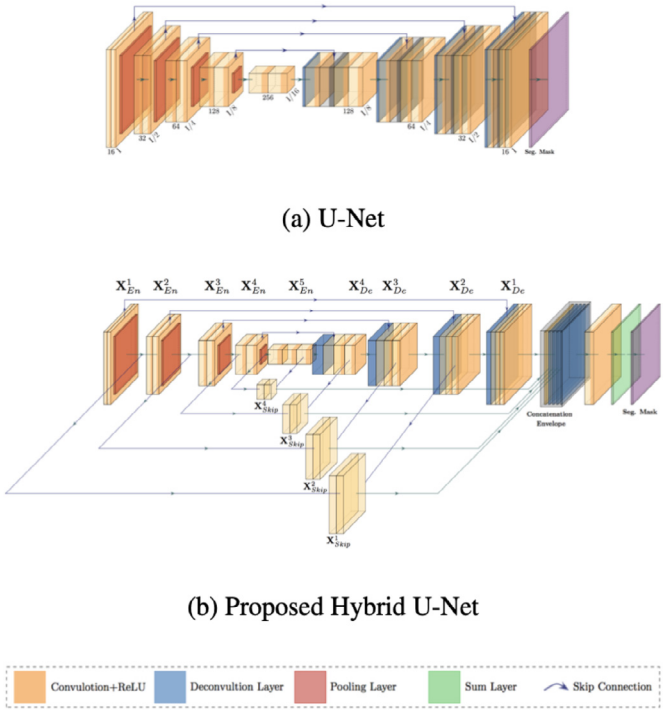

Fig. 3. Architecture of U-Net and Hybrid U-Net.

In addition to multi-scale U-Net variants, we also experiment with several other baseline models to show the effectiveness of U-Net family for this task; namely, ResNet50 (He et al., 2016), SegNet (Badrinarayanan et al., 2017), DeepLabv3+ (Chen, Zhu, Papandreou, Schroff, & Adam, 2018), and Swin-Unet (Cao et al., 2021). SegNet is a U-shaped encoder–decoder architecture that uses the encoder's pooling indices to upsample the decoder's feature maps. DeepLabv3+ attempts to combine the Atrous Spatial Pyramid Pooling (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2017) module with an encoder–decoder structure to better capture object boundaries. ASPP has shown to be very effective in encoding multi-scale contextual information; however, reconstructing refined high-resolution segmentation maps is the crucial part. Swin-Unet is a novel Transformer network with a U-Net-like structure, using two consecutive Swin transformer blocks in each decoder and encoder stage. Swin transformer (Liu et al., 2021) has shown to be computationally effective for semantic segmentation of high-resolution images and robust to large variations in the scale of visual entities. Table 3 summarizes models and number of their parameters.

Table 3
Number of parameters.

Method	U-Net	U-Net++	MACU-Net	Hybrid U-Net
Parameters (M)	1.2	2.3	5.2	3.5
Method	Swin-Unet	ResNet50	SegNet	DeepLabv3+
Parameters (M)	4.1	6.4	1.4	2.1

We also included two other variations of our proposed network. Hybrid U-Net V1 uses the architecture in Fig. 3(b), excluding the skip pathways from the encoder sub-network, by utilizing only the multiscale features from the decoder sub-network. Hybrid U-Net V2 utilizes deep supervision on the single-channel full-resolution feature maps from the skip pathways. In other words, we upsampled X^n_{Skip}s using single-channel transposed convolution and then deep supervised the network by optimizing the binary cross-entropy loss for each of the four semantic levels.

5.2. Metrics

We report a combination of segmentation and classification evaluation metrics for the complete evaluation of our proposed architecture. The Intersection over Union (IoU) or Jaccard index is the area of overlap of the ground truth mask and the predicted mask divided by the size of their union. This shows how well a semantic segmentation model performs in both the localization and classification tasks.

Let n_{ij} be the number of pixels of class i, predicted to belong to class j and let $N_i = \sum_j^{N_{classes}} n_{ij}$ be the total number of pixels of class i. Eq. (3) calculates the unweighted mean IoU (mIoU) over different classes. To consider the class imbalance, we also used the Frequency Weighed IoU (FWIoU), which is the IoU weighted by its sample size and then averaged over the classes.

$$mIoU = \frac{1}{N_{classes}} \sum_i \frac{n_{ii}}{\sum_j (n_{ij} + n_{ji}) - n_{ii}} \tag{3}$$

$$FWIoU = \frac{1}{\sum_k N_k} \sum_i \frac{N_i \; n_{ii}}{\sum_j (n_{ij} + n_{ji}) - n_{ii}} \tag{4}$$

FWIoU is more biased than mIoU toward models that perform well on the minority class at the expense of the majority class, a property that is quite interesting when dealing with an imbalanced dataset.

The Mean Dice Similarity Coefficient (MDC) is another widely used evaluation metric for image segmentation tasks. Similar to the IoU, it takes into account the area of overlap of the ground truth and the predicted masks but gives more weight to true positives. It divides twice the area of overlap by the total size of all the class samples.

$$MDC = \frac{1}{N_{classes}} \sum_i \frac{2 \; n_{ii}}{\sum_j (n_{ij} + n_{ji})} \tag{5}$$

We used another segmentation metric, the overall accuracy, which is calculated by dividing the number of correct predictions by the sample size. It is worth mentioning that in presence of class imbalance, high pixel accuracy does not always imply superior segmentation ability. This is due to the fact that if a class dominates the image, while some other classes make up only a small portion of the image, predicting the majority class for all pixels still leads to high accuracy. We also used the area under the Receiver Operating Characteristic (ROC) curve to assess the classification performance.

5.3. Training details

We used Keras as the deep learning framework and an Amazon Web Service (AWS) G4 instance with an NVIDIA T4 GPU with a 16 GB memory to train all the models. We randomly selected 60% of the images as the training set, 10% of the images as the validation set, and the remaining 30% as the test set. As we mentioned, there is no spatial overlap between covered areas. Using the Adam optimizer with a learning rate of 1e-4 and a cosine annealing decay, training and validation loss decreased smoothly to 0.19 and 0.22, respectively.

Table 4
Experiment results. Between baseline approaches, Swin-Unet performs the best in all of the metrics. Hybrid U-Net outperforms it in every metric, with 3.44%, 3.90%, and 1.11 improvements in mIoU, MDC, and area under the ROC curve, respectively. Between U-Net and its multiscale variants, raw U-Net achieves the best FWIoU and overall accuracy of 85.09% and 91.28%, respectively. MACU-Net, on the other hand, obtained the best mIoU, MDC, and area under the ROC curve of 60.63%, 71.61%, and 89.68, respectively. Between the proposed Hybrid U-Net's variations, the main architecture performs the best, with 1.31%, 4.85%, 4.49%, 1.70, and 1.10% improvements in FWIoU, mIoU, MDC, area under the ROC curve, and overall accuracy, respectively, compared with those of the best U-Net variant model.

Arch.	FWIoU	mIoU	MDC	ROC-AUC	OA
Swin-U	86.20	62.04	72.20	90.27	92.29
ResNet50 + FCN-8s	82.93	59.98	68.35	83.68	89.36
SegNet	81.57	45.16	47.46	74.56	90.32
DeepLab-V3	78.39	49.35	57.82	69.62	85.69
U-Net	85.09	58.43	68.01	88.59	91.28
U-Net++	72.32	48.26	53.11	70.17	75.04
MACU-Net	83.13	60.63	71.61	89.68	90.50
Hybrid U-net, V1	85.81	64.13	74.79	90.34	91.87
Hybrid U-net, V2	84.20	54.54	62.81	90.33	91.14
Hybrid U-net	**86.40**	**65.48**	**76.10**	**91.38**	**92.38**

6. Results

6.1. Damage detection accuracy

The results of the experiments with the different methods on the proposed dataset are shown in Table 4. The proposed Hybrid U-Net performed better than the other algorithms in all the quantitative evaluation indices. The main Hybrid U-Net achieved the area under ROC curve of 92.38 and FWIoU, mIoU, MDC, and accuracy of 86.40%, 65.48%, 76.10%, and 92.38%, respectively. Among baselines, Swin transformer-based model achieves the best results, outperforming U-Net and its previous multi-scale variations. Hybrid U-Net achieved better results, with 0.20%, 3.42%, 3.90%, 1.11, and 0.09% improvements over Swin-Unet in FWIoU, mIoU, MDC, AUC, and accuracy, respectively. The increase in FWIoU is less than the jump in mIoU because it is more biased toward the positive class, decreasing the effect of correct predictions for the negative class. ResNet50 backbone with FCN-8s resulted in the highest among other ResNet models, with mIoU and MDC similar to those of the raw U-Net. SegNet and DeepLabv3+ did not perform well, resulting in AUC of 74.56 and 69.62, respectively.

Among U-Net variants, raw U-Net and MACU-Net perform better than U-Net++. Compared with U-Net, Hybrid U-Net increased the mean dice coefficient, mIoU, and FWIoU by about 8%, 7%, and 1.3%, respectively. Hybrid U-Net's mIoU, MDC, and ROC-AUC also surpassed those of MACU-Net, which were the best among the baseline U-Net variants approaches, by about 5%, 4.5%, and 1.5%, respectively.

Considering the other variations of Hybrid U-Net, it can be seen that the first version (Hybrid U-Net V1) still outperformed U-Net and MACU-Net. This implies that even merely adding skip pathways from the decoder sub-network can already significantly improve the performance of raw U-Net which highlights the importance of the lost semantic information of the deeper layers of the decoder sub-network. Despite the satisfactory results of the first variation of Hybrid U-Net, the second variation did not outperform U-Net except in overall accuracy; but it still outperformed U-Net++ and in some cases, MACU-Net. One concern with applying deep supervision to all hidden layers is that it may interfere with the performance of the network (Lee, Xie, Gallagher, Zhang, & Tu, 2015), which happened with Hybrid U-Net V2.

6.2. Segmentation map visualization

To qualitatively compare the results of U-Net, MACU-Net, Hybrid U-Net V1, Swin-Unet, and Hybrid U-Net, we present the final feature map

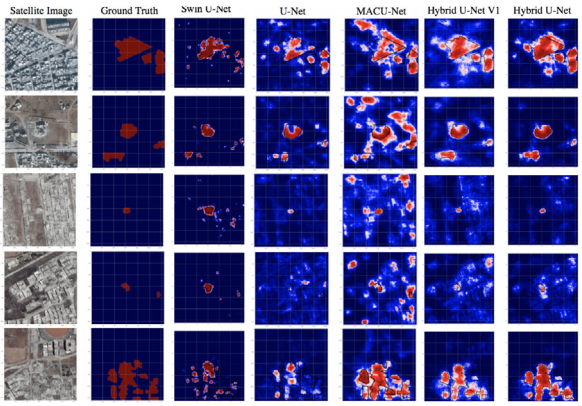

Fig. 4. Pixel scores heat maps.

scores in Fig. 4. Although U-Net performed decently where there was a small amount of destruction, as shown in the first and last rows, it tends to divide large levels of destruction into smaller ones. On the contrary, MACU-Net performed poorly in detecting low levels of destruction but performed better than U-Net on images showing a large amount of destruction. In addition, lines 1 and 2 of Fig. 4 clearly show that MACU-Net was more likely to mistake the bare soil areas as destruction, which makes it less suitable for segmenting images of rural areas.

The Hybrid U-Net results point out two significant advantages over other approaches: First, the contrast between the background and the destruction shows the network's certainty about its predictions. And second, its performance was satisfactory for both small and large destroyed areas. More specifically, it can detect large destroyed areas as a whole and distinguish adjacent destruction, which is displayed very well in line 5 of Fig. 4.

The best baseline model, Swin-Unet performs a decent job in detecting both classes more certainly, but boundaries are less accurate compared with the Hybrid U-Net, due to the fact that the crucial spatial information is still lost.

6.3. Robustness

As discussed in the previous experiment, the compared methods appeared to have performed differently based on the amount of area that is destroyed. We further investigated this idea by comparing the mIoU and the MDC of the different approaches based on the fraction of the destructions in each image. We obtained the fraction of the destroyed area in each image by dividing the size of the positive class, i.e., the destroyed area, by the size of the image. Fig. 5 shows the mIoU and the MDC of each image versus the fraction destroyed. As demonstrated, Hybrid U-Net and its variants achieved more than 15% and 25% improvements in their mIoUs and MDCs upon increasing the fraction of the destructions, whereas U-net, MACU-Net, and Swin-Unet had less than 7%, 3%, and %0.5 improvements in their mIoUs and 19%, 14%, and 5% improvements in their MDCs, respectively. For images with a very small fraction of destructions, Swin-Unet performs about 5% better than Hybrid U-Net in mIoU and MDC; however, Hybrid U-Net outperforms Swin-Unet in both mIoU and MDC for fractions larger than 5%.

7. Conclusion

In this paper, we investigated the performance of semantic segmentation architectures for detecting war-inflicted building destruction. We showed that even with a few high-resolution satellite images, U-Net and its variants result in accurate destruction detection. We also introduced a dataset of satellite images of Syria that we pixel-wise-annotated to mark the destroyed areas. We further proposed a

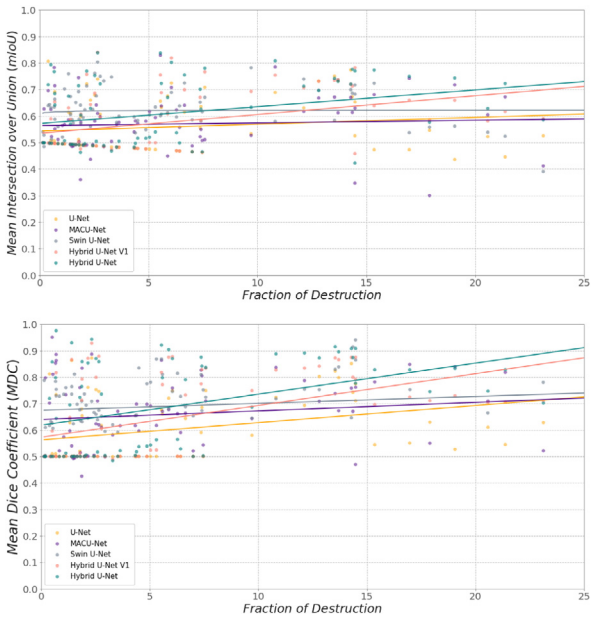

Fig. 5. The mIoU and MDC versus the fraction of the destructions. Fitted lines are obtained by minimizing the squared errors. All U-Net models achieve the mIoU of roughly 0.55 for low level of destructions. The Hybrid U-Net variations reach the mIoU of about 0.70 for larger ones, while the mIoU of the U-Net and MACU-Net increases to roughly 0.60. Swin-Unet is almost invariant to the destruction size, staying at about 0.62 for any fraction of destruction. The MDC starts at about 0.60 and increases to about 0.85 for Hybrid U-Net variations and 0.75 for U-Net and MACU-Net. Swin-Unet starts at 0.68 and increases by 0.05 for the larger fraction of destruction.

symmetrical multiscale skip-connected architecture based on U-Net for the semantic segmentation of high-resolution satellite images. The multiple quantitative and qualitative experiments on the dataset confirmed the superior performance of Hybrid U-Net over baseline approaches. In addition, due to the high variations of the amount of destruction shown in each image, we tested the performance of the proposed and benchmark networks versus the size of the positive samples in each image. Our findings showed that apart from the superior performance of Hybrid U-Net on images with low levels of destruction, it performed significantly better than the benchmark approaches in its segmentation of images with a large fraction of destruction.

CRediT authorship contribution statement

Shima Nabiee: Methodology, Data curation, Software, Investigation, Visualization, Writing. **Matthew Harding:** Supervision. **Jonathan Hersh:** Conceptualization, Data curation. **Nader Bagherzadeh:** Project administration, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: University of California, Irvine Chapman University

References

Ahmed, I., Ahmad, M., Khan, F. A., & Asif, M. (2020). Comparison of deep-learning-based segmentation models: Using top view person images. *IEEE Access*, *8*, 136361–136373.

Avtar, R., Kouser, A., Kumar, A., Singh, D., Misra, P., Gupta, A., et al. (2021). Remote sensing for international peace and security: Its role and implications. *Remote Sensing*, *13*(3), 439.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2481–2495.

Braun, A. (2019). *Radar satellite imagery for humanitarian response* (Ph.D. thesis), Germany: Universit of Tübingen, Tübingen.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537.

Chen, L. -C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Chen, L. -C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision* (pp. 801–818).

Fisher, S., Matovic, V., Ludin, J., Abdi, D. I., Walker, B. A., Smith, R., et al. (2000). *Working with conflict 2: Skills and strategies for action*. Zed books.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., et al. (2020). UNet 3+: A full-scale connected UNet for medical image segmentation. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 1055–1059).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Jabri, V. (1996). *Discourses on violence: Conflict analysis reconsidered*. Manchester University Press.

Kishi, N. (2021). Satellite data and crowdsourcing. *Space Policy*, *56*, Article 101423.

Knoth, C., Slimani, S., Appel, M., & Pebesma, E. (2018). Combining automatic and manual image analysis in a web-mapping application for collaborative conflict damage assessment. *Applied Geography*, *97*, 25–34.

Korznikov, K. A., Kislov, D. E., Altman, J., Doležal, J., Vozmishcheva, A. S., & Krestov, P. V. (2021). Using U-Net-like deep convolutional neural networks for precise tree recognition in very high resolution RGB (red, green, blue) satellite images. *Forests*, *12*(1), 66.

Lee, K. S., Chen, H. L., Ng, Y. S., Maul, T., Gibbins, C., Ting, K. -N., et al. (2022). U-Net skip-connection architectures for the automated counting of microplastics. *Neural Computing and Applications*, 1–15.

Lee, C. -Y., Xie, S., Gallagher, P., Zhang, Z., & Tu, Z. (2015). Deeply-supervised nets. In *Artificial intelligence and statistics* (pp. 562–570). PMLR.

Li, R., Duan, C., Zheng, S., Zhang, C., & Atkinson, P. M. (2021). MACU-Net for semantic segmentation of fine-resolution remotely sensed images. *IEEE Geoscience and Remote Sensing Letters*.

Lin, T. -Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Mao, X. -J., Shen, C., & Yang, Y. -B. (2016). Image restoration using convolutional auto-encoders with symmetric skip connections. arXiv preprint arXiv:1606.08921.

Marx, A. J., & Loboda, T. V. (2013). Landsat-based early warning system to detect the destruction of villages in Darfur, Sudan. *Remote Sensing of Environment*, *136*, 126–134.

Mason, S., & Rychard, S. (2005). *Conflict analysis tools-tip sheet*. ISN-ETH Zurich.

Mueller, H., Groeger, A., Hersh, J., Matranga, A., & Serrat, J. (2021). Monitoring war destruction from space using machine learning. *Proceedings of the National Academy of Sciences*, *118*(23).

Nabiee, S., Hersh, J., Harding, M., & Bagherzadeh, N. (2021). Syria civil war destructions dataset for hybrid-U-Net. http://dx.doi.org/10.5281/zenodo.1234.

Pagot, E., & Pesaresi, M. (2008). Systematic study of the urban postconflict change classification performance using spectral and structural features in a support vector machine. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *1*(2), 120–128.

Phan, T. -D. -T., Kim, S. -H., Yang, H. -J., Lee, G. -S., et al. (2021). Skin lesion segmentation by U-Net with adaptive skip connection and structural awareness. *Applied Sciences*, *11*(10), 4528.

Quinn, J. A., Nyhan, M. M., Navarro, C., Coluccia, D., Bromley, L., & Luengo-Oroz, M. (2018). Humanitarian applications of machine learning with remote-sensing data: Review and case study in refugee settlement mapping. *Philosophical Transactions of the Royal Society of London A (Mathematical and Physical Sciences)*, *376*(2128), Article 20170363.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-Net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*.

Sulik, J. J., & Edwards, S. (2010). Feature extraction for Darfur: Geospatial applications in the documentation of human rights abuses. *International Journal of Remote Sensing*, *31*(10), 2521–2533.

Tarasiewicz, T., Nalepa, J., & Kawulok, M. (2020). Skinny: A lightweight U-Net for skin detection and segmentation. In *2020 IEEE international conference on image processing* (pp. 2386–2390). IEEE.

UK Government's Stabilisation Unit (2017). *Joint analysis of conflict and stability: Guidance note*. London: Stabilisation Unit.

Witmer, F. D. W. (2015). Remote sensing of violent conflict: Eyes from above. *International Journal of Remote Sensing*, *36*(9), 2326–2352.

Wolfinbarger, S., & Wyndham, J. (2011). Remote visual evidence of displacement. *Forced Migration Review*, (38), 20.

Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 1395–1403).

Yuan, M., Liu, Z., & Wang, F. (2019). Using the wide-range attention U-Net for road segmentation. *Remote Sensing Letters*, *10*(5), 506–515.

Zhang, J., Jin, Y., Xu, J., Xu, X., & Zhang, Y. (2018). MdU-Net: Multi-scale densely connected U-Net for biomedical image segmentation. arXiv preprint arXiv:1812.00352.

Zhang, Z., Liu, Q., & Wang, Y. (2018). Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, *15*(5), 749–753.

Zhang, Y., Wu, J., Chen, W., Liu, Y., Lyu, J., Shi, H., et al. (2019). Fully automatic white matter hyperintensity segmentation using U-Net and skip connection. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society* (pp. 974–977). IEEE.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested U-Net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 3–11). Springer.