Remote Wildfire Detection using Multispectral Satellite Imagery and Vision Transformers [*]

Ryan Rad R.RAD@NORTHEASTERN.EDU
Khoury College of Computer Science, Northeastern University, Vancouver, BC, Canada

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Wildfires pose a significant and recurring challenge in North America, impacting both human and natural environments. The size and severity of wildfires in the region have been increasing in recent years, making it a pressing concern for communities, ecosystems, and the economy. The accurate and timely detection of active wildfires in remote areas is crucial for effective wildfire management and mitigation efforts. In this research paper, we propose a robust approach for detecting active wildfires using multispectral satellite imagery by leveraging vision transformers and a vast repository of landsat-8 satellite data with a 30m spatial resolution in North America. Our methodology involves experimenting with vision transformers and deep convolutional neural networks for wildfire detection in multispectral satellite images. We compare the capabilities of these two architecture families in detecting wildfires within the multispectral satellite imagery. Furthermore, we propose a novel u-shape vision transformer that effectively captures spatial dependencies and learns meaningful representations from multispectral images, enabling precise discrimination between wildfire and non-wildfire regions. To evaluate the performance of our approach, we conducted experiments on a comprehensive dataset of wildfire incidents. The results demonstrate the effectiveness of the proposed method in accurately detecting active wildfires with an *Dice Score or F*1 of %90.05 and *Recall* of %89.61 . Overall, our research presents a promising approach for leveraging vision transformers for multispectral satellite imagery to detect remote wildfires.

Keywords: Vision Transformer, Wildfire Detection; Landsat-8; Multispectral Imaging; Remote Sensing; Satellite Imagery

1. Introduction

Wildfires have become a growing concern in North America, posing significant threats to both human lives and natural environments. Over the past years, the region has witnessed an alarming increase in the size and severity of wildfires, necessitating the development of effective detection and mitigation strategies. Timely and accurate identification of active wildfires in remote areas is crucial for minimizing the damage caused by these destructive events and enabling prompt firefighting and evacuation efforts.

Traditional methods of wildfire detection heavily rely on ground-based observations and weather monitoring systems. However, these approaches often face limitations in terms of coverage, scalability, and real-time detection capabilities. To overcome these challenges, the

[*] Thanks to Google Cloud Research Credit for providing us with access to Google's high-performance computing capabilities for this research project.

use of remote sensing technologies, particularly satellite imagery, has emerged as a powerful tool for wildfire detection and monitoring.

Multispectral satellite imagery, such as the Landsat-8 dataset with a spatial resolution of 30 meters, provides a comprehensive view of the Earth's surface by capturing data across multiple spectral bands. This rich source of information offers valuable insights into the characteristics of wildfire-affected areas, including changes in vegetation, heat signatures, and smoke plumes. Leveraging this wealth of multispectral data, along with advanced machine learning techniques, holds great potential for enhancing the accuracy and efficiency of wildfire detection systems.

In recent years, deep learning-based approaches, such as convolutional neural networks (CNNs), have shown remarkable success in various computer vision tasks, including object recognition and image classification. These methods have also been applied to wildfire detection, leveraging satellite imagery to automatically identify fire-affected regions. However, CNN-based models have limitations in capturing long-range dependencies and understanding the complex spatial relationships present in multispectral data.

To address these limitations, we propose the integration of vision transformers, a recent advancement in deep learning, into the task of active wildfire detection. Vision transformers have demonstrated exceptional performance in various computer vision tasks, surpassing the capabilities of CNNs in capturing global dependencies and modeling image context. By applying vision transformers to multispectral satellite imagery, we aim to leverage their strengths in learning meaningful representations and spatial dependencies, enabling more precise discrimination between wildfire and non-wildfire regions.

In this research paper, we present a robust approach for active wildfire detection using multispectral satellite imagery and vision transformers. Our methodology involves the development of deep U-shape Vision Transformer that is trained on a vast repository of Landsat-8 satellite data in North America. We compare the performance of alternative architecture families in detecting wildfires within multispectral satellite images, evaluating their accuracy and efficiency.

To validate the effectiveness of our approach, we conduct experiments on a comprehensive dataset of wildfire incidents, assessing the detection performance through metrics such as Dice Score or F1 and Recall. We used a dataset for active fire detection consisting of over $150,000$ image patches extracted from Landsat-8 satellite images captured worldwide in August and September 2020. We utilized five specific spectral bands: SWIR1, SWIR2, HCHO (formaldehyde), landcover maps, and the evaporation index as depicted in Figure 1. In selecting the 5 input bands, our criteria aimed to maximize the diversity of spectral information captured by the chosen bands while mitigating potential multicollinearity issues, thus enhancing the model's capacity to discriminate between wildfire and non-wildfire regions. The results demonstrate the promising capabilities of our proposed method in accurately detecting remote wildfires, thereby contributing to the advancement of wildfire management and mitigation efforts.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work in the field of deep convolutional neural networks and vision transformers. Section 3 describes the methodology, including the architecture design of our deep U-shape Vision Transformer. Section 4 presents the experimental setup, including the dataset used as well as results and analysis, followed by concluding remarks in Section 5.

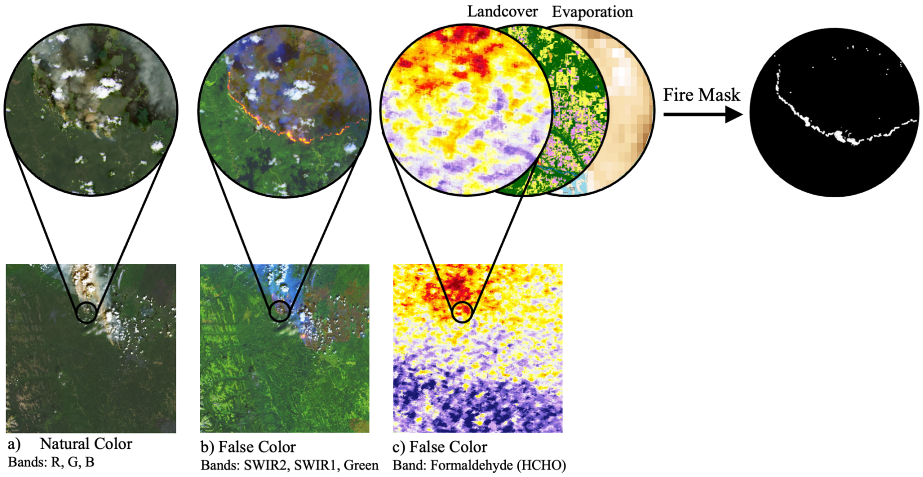


Figure 1: Sample spectral bands visualization above Klua Lake, BC (June 2023).

2. Related Work

- *Deep Convolutional Neural Networks (DCNN):* DCNN-based methods have been widely used for image segmentation or pixel-wise classification, with U-Net Ronneberger et al. (2015) being a popular choice due to its simplicity and performance. Several variants of U-Net Rad et al. (2019a,b); Iglovikov and Shvets (2018); Rad et al. (2018c) have also been proposed to further improve the performance. Global contextual information is widely recognized as advantageous for semantic segmentation Luo et al. (2016); Zhao et al. (2017); Chen et al. (2017). *PSPNet* Zhao et al. (2017) introduced a pyramid pooling module that applies pooling operations at various scales, while *DeepLabv3* Chen et al. (2017) proposed parallel Atrous convolution with different rates to integrate global context. However, the pooling operation with striding in Zhao et al. (2017) may result in information loss at object boundaries, and the use of dilated convolution with a large dilation rate in Chen et al. (2017) can give rise to the "grinding" problem. CNN-based methods have achieved remarkable success in this field, thanks to their powerful representation capabilities.

- *Vision Transformers:* Transformers Vaswani et al. (2017) initially gained prominence in natural language processing (NLP) tasks, where they achieved state-of-the-art performance. Building on this success, they have now made their way into the field of computer vision. Vision transformers have the unique ability to capture global dependencies and long-range interactions, making them valuable for tasks like image recognition, object detection, and semantic segmentation. Unlike convolutional neural networks (CNNs), which typically require fixed input dimensions, vision transformers offer a flexible and scalable architecture that can handle images of varying sizes. Vision Transformers (ViT) Dosovitskiy et al. (2020) were introduced to handle image recognition tasks, but they require pre-training on large datasets. To address this, approaches like Deit Touvron et al. (2020) have been proposed to improve the training of ViT. A notable example is Swin TransformerSwin Transformer Liu et al. (2022), an efficient hierarchical vision Transformer, for various vision tasks, including image classification, object detection, and semantic segmentation. One notable example of this family of models is Cao et al. (2022).

- *Self-attention/Transformer in conjunction with DCNNs:* Researchers have also explored the combination of self-attention mechanisms, typically found in Transformers, with CNNs to enhance network performance. Some approaches integrate self-attention with CNN-based U-shaped architectures for medical image segmentation. Other studies focus on combining Transformers and CNNs to improve segmentation capabilities, particularly in multi-modal brain tumor segmentation and 3D medical image segmentation. Two notable examples of this family of models Chen et al. (2021) and Zhang et al. (2021).

3. Methodology

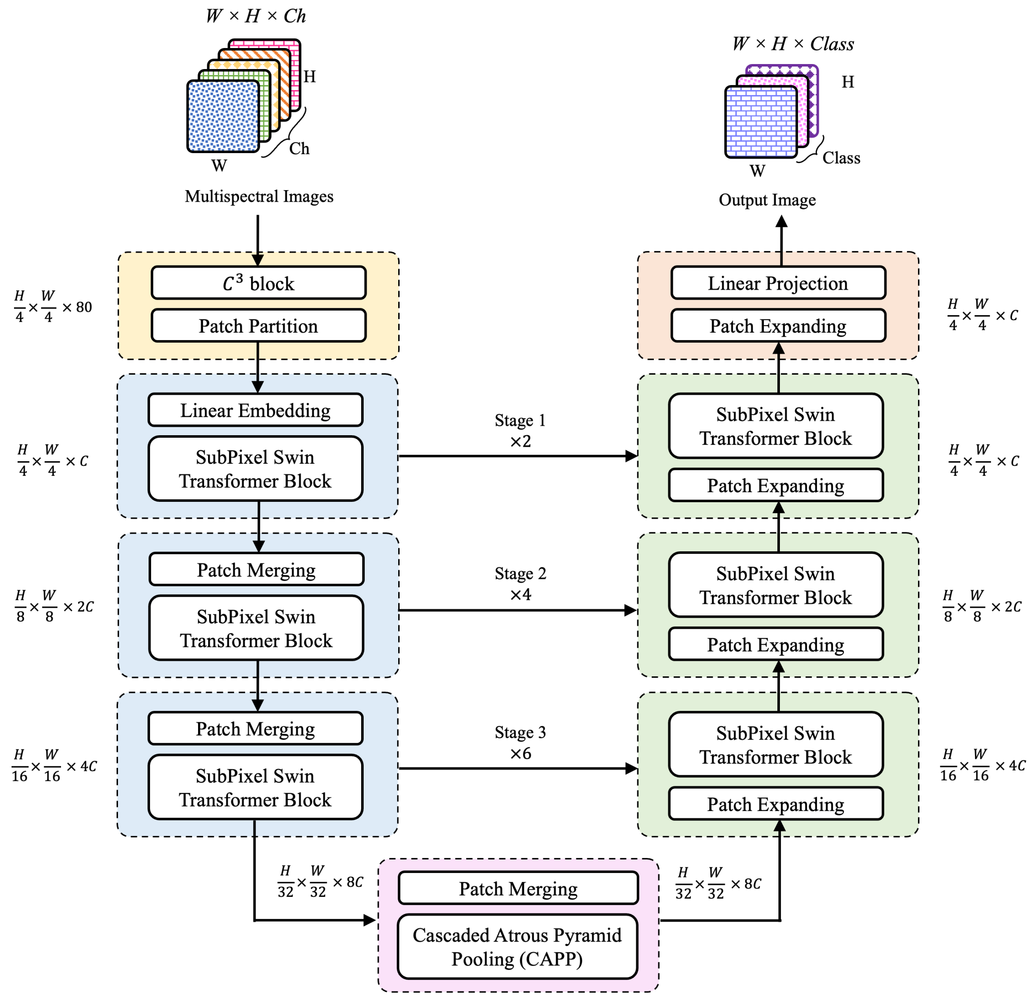
3.1. Overall Architecture

Building upon the remarkable success of the Swin Transformer Liu et al. (2022), we present a novel U-shaped Encoder-Decoder architecture called SubPixel-Swin (SP-Swin) Unet, specifically designed for multispectral images. Our approach leverages the advantages of the Swin Transformer and combines them with the U-shaped architectural design. In SP-Swin Unet, both the encoder and decoder components are constructed using SP-Swin Transformer blocks, resulting in a Transformer-based U-shaped architecture tailored for multispectral image analysis.

Figure 2 provides an overview of the architecture of the SP-Swin Transformer, highlighting the best performing version. The architecture shares a similar structure for stages 1 to 4 with the Swin Transformer, but it incorporates two key differences. First, Stage 0 is introduced to enhance the model's ability to learn from multi-spectral images. Second, the Shifted window partitioning used in the original Swin Transformer is replaced with Sub-Pixel Window Partitioning in the SP Swin Transformer. The patch partitioning module splits an input multi-spectral image into non-overlapping patches, like ViT and Swin. In the SubPixel Swin (SP Swin) Transformer, individual patches are treated as ”tokens” with their features constructed by concatenating the raw values from five multi-spectral bands. Assuming a patch size of 4×4, each patch's feature dimension is calculated as $4 \times 4 \times 5 = 80$. To project these raw-valued features into an arbitrary dimension represented as C, a linear embedding layer is applied. The resulting patch tokens, along with their embedded features, are processed through multiple Swin Transformer blocks, which have modified self-attention computations.

The proposed SP-Swin Transformer architecture, applies several Transformer blocks with modified self-attention computation to patch tokens. To create a hierarchical representation, patch merging layers are used to reduce the number of tokens as the network goes deeper in a similar fashion as Swin Transformer. The first patch merging layer concatenates the features of neighboring 2×2 patches and applies a linear layer to the concatenated features. This reduces the number of tokens by a factor of 4 downsampling of resolution) and sets the output dimension to $2C$. More Transformer blocks are then applied for feature transformation, keeping the resolution at $\frac{H}{8} \times \frac{W}{8}$. This process is repeated twice for ”Stage 3” and ”Stage 4,” resulting in output resolutions of $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, respectively.

By utilizing these stages, the Swin Transformer generates a hierarchical representation with the same feature map resolutions as traditional convolutional networks like VGG and ResNet. This allows the proposed architecture to easily replace the backbone networks in existing methods for various vision tasks.



$W \times H \times Ch$
H
Ch
W
Multispectral Images
$W \times H \times Class$
H
Class
W
Output Image
C^3 block
Patch Partition
$\frac{H}{4} \times \frac{W}{4} \times 80$
Linear Projection
Patch Expanding
$\frac{H}{4} \times \frac{W}{4} \times C$
Linear Embedding
SubPixel Swin Transformer Block
$\frac{H}{4} \times \frac{W}{4} \times C$
Stage 1 ×2
SubPixel Swin Transformer Block
Patch Expanding
$\frac{H}{4} \times \frac{W}{4} \times C$
Patch Merging
SubPixel Swin Transformer Block
$\frac{H}{8} \times \frac{W}{8} \times 2C$
Stage 2 ×4
SubPixel Swin Transformer Block
Patch Expanding
$\frac{H}{8} \times \frac{W}{8} \times 2C$
Patch Merging
SubPixel Swin Transformer Block
$\frac{H}{16} \times \frac{W}{16} \times 4C$
Stage 3 ×6
SubPixel Swin Transformer Block
Patch Expanding
$\frac{H}{16} \times \frac{W}{16} \times 4C$
$\frac{H}{32} \times \frac{W}{32} \times 8C$
Patch Merging
Cascaded Atrous Pyramid Pooling (CAPP)
$\frac{H}{32} \times \frac{W}{32} \times 8C$

Figure 2: Architecture of the proposed SubPixel-Swin (SP-Swin) Unet.

3.2. Compact-Contextualize-Calibrate (C^3) Block

Recently, there has been a surge in research focusing on Vision Transformers, showcasing their effectiveness across a broad spectrum of computer vision tasks. Here, we investigate a different application of Vision Transformers, which involves leveraging contextual information to learn from multispectral images.

The *Compact-Contextualize-Calibrate* (C^3) block is a computational flow designed to enhance the learning of features from multiple input bands, particularly when there is limited correlation among those bands. This flow enables more effective feature extraction and representation in scenarios where the bands exhibit less mutual correlation by recalibrating the input spatially and channel-wise.

Let F_{transf} be a convolutional operation and $K = [k_1, k_2, ..., k_n]$ a set of learned kernels. Then, transformed outputs are obtained by convolving the kernels over the input X: $U_i = K_i * X$, $X \in \mathbb{R}^{W \times H \times C}$, $U \in \mathbb{R}^{W' \times H' \times Ch'}$. The C^3 unit can be employed to processes the input X to generate context-aware X' prior to performing the F_{transf}. The input X goes through three layers that are depicted in Fig. 3 and explained next.

- *Compact:* In order to identify key locations of the input, a spatial-wise compaction is performed over all the input channels. This process generates a spatial-wise statistic $Z \in \mathbb{R}^{W \times H \times 1}$ by compacting X, such the i, j-th element of Z is calculated by:

$$Z_{i,j} = F_{compt}(X) = \frac{1}{N} \sum_{n=1}^{N} X_n(i, j) \tag{1}$$

- *Contextualize:* To further capture the contextual information, another step is taken by applying a flexible operation. This operation must be able to learn contextual relationships. To achieve this, a two-step down- up-sampling operation of factor r ($r = 4 - 6$ are empirically shown to be the most effective) is performed:

$$V = F_{contxt}(Z) = W_2 \delta(W_1 Z) \tag{2}$$

  Strided convolution is utilized as W_1 for down-sampling and sub-pixel convolution as W_2 for up-sampling and δ is referred to the ReLu activation function.

- *Calibrate:* Finally, we direct the attention of the feature extraction to the key locations of the input since not two pixels are worth the same. The (i, j)-th element of X_c is calibrated by:

$$X'_c(i, j) = F_{cal}(X_c, V) = X_c(i, j) \cdot \sigma(V(i, j)) \tag{3}$$

  Here σ is referred to as the sigmoid activation function to bring V into the range of $(0, 1)$.

It should be noted that C^3 blocks are designed to be lightweight, ensuring that they do not impose a significant increase in model complexity or computational costs. The C^3 block differs from the SE (Squeeze-and-Excitation) block introduced in Hu et al. (2018) in two key aspects. Firstly, the C^3 block operates on both spatial and channel dimensions. Secondly, while the SE block aims to refine learned features after CNN-based feature extraction, the C^3 block contextualizes input features prior to feature transformation within a transformer framework.

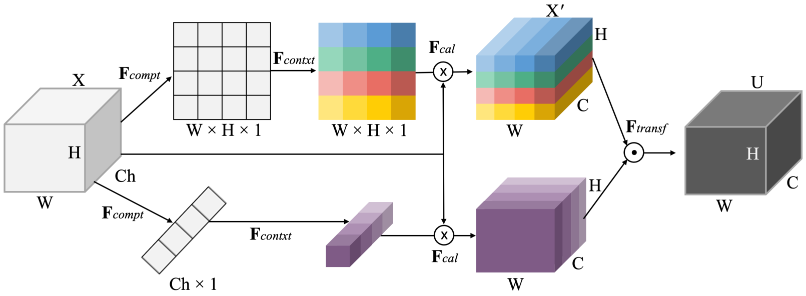


Figure 3: Structure of the Compact-Contextualize-Calibrate (C^3) Block

3.3. Cascaded Atrous Pyramid Pooling (CAPP)

A module involving depth-wise dilated convolution is introduced prior to stage 1 transformer block in the Swin Transformer for multispectral images to effectively capture spatial context, extract relevant features, and optimize efficiency in processing the high-dimensional multispectral data.

Multispectral images have spatially correlated information across bands. Depth-wise dilated convolution captures local and contextual information, leveraging the spatial context and extracting comprehensive representations. Depth-wise dilated convolution captures multi-scale patterns and structures by extracting features at different dilation rates. This helps represent the diverse and complex characteristics of multispectral images. Finally, Depth-wise dilated convolution reduces the number of parameters compared to traditional convolutional layers, making it more efficient for processing high-dimensional multispectral data.

Different approaches have been invented to capture global contextual information. For example, *PSPNet* Zhao et al. (2017) proposed a pyramid pooling module by applying pooling operations at different scales, while *DeepLabv3* Chen et al. (2017) proposed parallel Atrous convolution with different rates to incorporate global context. The pooling operation with striding in Zhao et al. (2017) could lead to the information loss at object boundaries and applying dilated convolution with a large dilation rate in Chen et al. (2017) could raise the 'grinding' problem.

Dilated convolution Yu and Koltun (2016) has been increasingly utilized in the architecture of several DCNNs Rad et al. (2018b,a); Yu and Koltun (2016); Yu et al. (2017); Wang et al. (2017); Chen et al. (2017) to enlarge the receptive field without introducing additional parameters to the network. In the dilated convolution, a kernel K with size $s \times s$ and the dilation rate of r only visits its input signal F at every r^{th} location of each dimension, as defined in Eq. 1:

$$(F *_r K)(x,y) = \sum_{m=-t}^{t} \sum_{n=-t}^{t} K(m,n)F(x-r.m, y-r.n)$$
$$\text{where} \quad t = (s-1)/2 \tag{4}$$

According to Eq. 1, from a $s_d \times s_d$ dilated neighbourhood region, where $s_d = (r-1).(s-1)+s$, only $s \times s$ pixels contribute to the computation of the response. All the $s \times s$

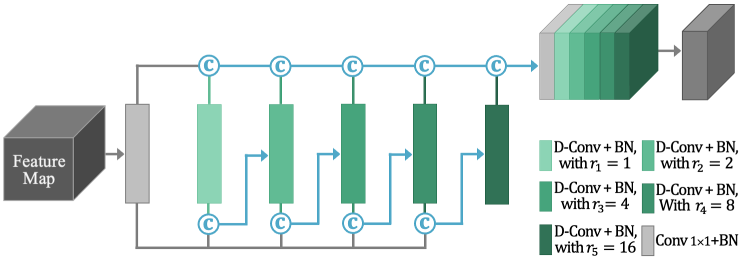

Feature Map
c
D-Conv + BN, with $r_1 = 1$
D-Conv + BN, with $r_2 = 2$
D-Conv + BN, with $r_3 = 4$
D-Conv + BN, With $r_4 = 8$
D-Conv + BN, with $r_5 = 16$
Conv 1×1+BN

Figure 4: Structure of the Cascaded Atrous Pyramid Pooling (CAPP).

contributing pixels have the same distance of $r - 1$ pixels from each other and the centroid. For example, in a 3×3 dilated kernel with $r = 4$, only 9 pixels (out of the 81) contribute to the kernel response. This under-utilization of information amounts to approximately $\sim 89\%$. Furthermore, as the dilation rate increases, the correlation between pixels gradually diminishes, resulting in a reduction in the number of valid weights. Eventually, the 3×3 kernel behaves similar to a 1×1 kernel, as noted in Chen et al. (2017). These problems motivate the proposal of CAPP, a simple yet effective solution to address the root cause of the grinding problem.

The CAPP module, as illustrated in Fig. 4, pursues two primary objectives. First, it eliminates the grinding issue by giving every single pixel in the dilated neighborhood a role to participate in the computation of the kernel response. Second, it further enlarges the receptive field to facilitate analyzing the complex structure of an image. In CAPP, a large dilation rate of 2^j is backed up by smaller dilation rates of 2^i where $j > i \geq 0$ with skip connections. Particularly, the dilation rate is increased to 2^i at the $(i+1)^{th}$ level of the pyramid. In cascaded dilated convolution, not only does each pixel matter but also its contribution is somewhat proportional to its distance from the central pixel. Cascaded structure (i.e., instead of parallel) and concatenation with core features are two major differences between CAPP and ASPP in Chen et al. (2017).

3.3.1. SP-Swin Transformer block

Different from the conventional multi-head self-attention (MSA) module, swin transformer block Liu et al. (2022) is constructed based on shifted windows. In Figure 5, two consecutive SP-Swin transformer blocks are presented. Each SP-Swin transformer block is composed of LayerNorm (LN) layer, multi-head self-attention module, residual connection and 2-layer MLP with GELU non-linearity. The based multi-head self-attention (W-MSA) module and the Sub-Pixel multi-head self-attention (SP-MSA) module are applied in the two successive transformer blocks, respectively.

3.3.2. Sub-Pixel Window based Self-Attention

The use of sub-pixel shuffling window-based self-attention in the Swin Transformer instead of a shifted window can be justified for several reasons:

- *Reduced Information Loss:* The sub-pixel shuffling window allows for a more precise alignment of patches during the self-attention process. Unlike the shifted window approach, which shifts the patches by a fixed stride, the sub-pixel shuffling window

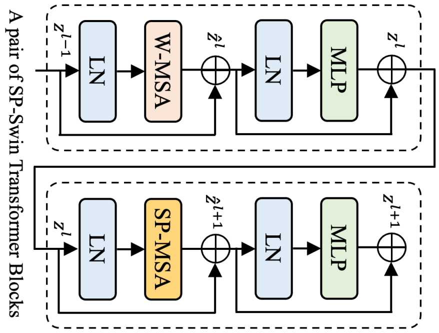


Figure 5: Two consecutive SP-Swin transformer blocks.

rearranges the patches in a way that minimizes information loss, as illustrated in Figure 6. This ensures that the attention mechanism can capture fine-grained details and preserve spatial information more effectively.

- *Enhanced Local Context:* By shuffling the patches using sub-pixel shuffling, the self-attention mechanism can capture local context more accurately. This is because neighboring patches, which contain related information, are placed closer to each other, allowing the attention mechanism to better capture the dependencies and relationships within the local context. The shifted window approach may introduce misalignment and reduce the ability to capture precise local relationships.

- *Improved Global Context:* The sub-pixel shuffling window also facilitates the capture of the global context in the self-attention mechanism. As the patches are rearranged, the attention mechanism can effectively attend to patches that are spatially distant but semantically related, as illustrated in Figure 6. This enables the model to capture long-range dependencies and incorporate global context information into the representation learning process.

- *Better Integration with Hierarchical Structure:* The sub-pixel shuffling window aligns well with the hierarchical structure of the Swin Transformer. As the network progresses through different stages, the patches are merged and the resolution decreases. The use of sub-pixel shuffling allows for a consistent alignment of patches across different stages, maintaining the coherence of attention patterns and facilitating information flow between different levels of the hierarchy.

Overall, the sub-pixel shuffling window-based self-attention in the Swin Transformer offers improved information preservation, enhanced local and global context modeling, and better integration with the hierarchical structure of the network. These advantages justify its use over the shifted window approach.

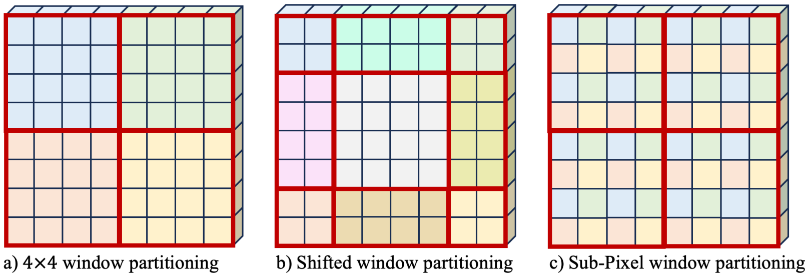

a) 4×4 window partitioning
b) Shifted window partitioning
c) Sub-Pixel window partitioning

Figure 6: Shifted window partitioning vs Sub-Pixel window partitioning.

4. Experimental Results

4.1. Experimental Setup

4.1.1. Dataset

We used a dataset for active fire detection consisting of over 150,000 image patches extracted from Landsat-8 satellite images captured worldwide in August and September 2020 from de Almeida Pereira et al. (2021). The dataset is divided into two parts: the first part includes 10-band spectral images with outputs from three well-known handcrafted algorithms for active fire detection, while the second part contains manually annotated masks. We focused on North America and obtained the ground truth labels by using a voting mechanism based on the outputs of the three algorithms. For this work, we utilized five specific spectral bands: SWIR1, SWIR2, HCHO (formaldehyde), landcover maps, and the evaporation index. These bands were selected to capture different aspects of the target environment and enable effective training of our model.

4.1.2. Implementation Details

The proposed *SP-Swin-UNet* model is implemented based on Python 3.8 and *Pytorch*1.8.0. In order to enhance the diversity of the training data, a comprehensive range of data augmentation techniques, including flips, scaling, translation, brightness, Gaussian noise, synthetic clouds, and rotations, are applied to all training patches. The input image size and patch size are set as 224 × 224 and 64, respectively. We train our model on two NVIDIA Tesla T4 with a total of 32GB memory on Google Cloud in northamerica-northeast2-Toronto region. During the training period, the SGD optimizer with momentum 0.9 for 100 epochs using a cosine decay learning rate scheduler and 20 epochs of linear warm-up.

4.2. Quantitative Results

4.2.1. Wildfire detection performance

To highlight the effectiveness of the proposed model, several widely adopted architectures such as Baseline *UNet* Ronneberger et al. (2015) and TernausNet Iglovikov and Shvets (2018) (winner of the Carvana challenge), PSPNet Zhao et al. (2017), DeepLab V3 Chen et al. (2017), Blast-Net Rad et al. (2019a), Trans UNet Chen et al. (2021), and Swin UNet Cao et al. (2022) are considered. Table 1 compares the performance of the proposed *Sub-Pixel Swin UNet* model to that of other models considered in this study. Both the

Table 1: Fire segmentation performance comparison (*Dice Score or F*1 *%*). The best results are in bold **black** and the second-best ones are in **teal**.

Models	Size	Dice (median)	Dice (stdev)	Recall	EC (kWh)	CO_2e (lbs)
UNet Ronneberger et al. (2015)	**6m**	79.98	**2.7**	78.25	**16.5**	**15.7**
TernausNet Iglovikov and Shvets (2018)	**10m**	82.12	5.1	79.61	24.5	23.4
PSPNet Zhao et al. (2017)	35m	84.69	4.9	82.92	21.3	20.3
DeepLab V3 Chen et al. (2017)	40m	85.38	4.8	83.54	20.8	19.8
Blast-Net Rad et al. (2019a)	25m	85.97	4.7	84.38	19.7	18.8
Trans UNet Chen et al. (2021)	42m	86.48	4.5	85.63	25.3	24.1
Swin UNet Cao et al. (2022)	27m	87.85	4.2	**88.79**	24.6	23.5
SP-Swin UNet w/o C^3	14m	**89.11**	3.5	88.53	18.9	18.0
SP-Swin UNe w/o CAPP	12m	88.27	4.1	87.37	19.7	18.8
SP-Swin UNet	15m	**90.05**	**2.9**	**89.61**	**18.2**	**17.4**

Table 2: Pairwise comparisons between models using the *Tukey's HSD* test.

Model 1	Model 2	p-value (adj)	Reject
SP-Swin UNet	UNet Ronneberger et al. (2015)	0.0000	True
SP-Swin UNet	TernausNet Iglovikov and Shvets (2018)	0.0000	True
SP-Swin UNet	PSPNet Zhao et al. (2017)	0.0010	True
SP-Swin UNet	DeepLab V3 Chen et al. (2017)	0.0019	True
SP-Swin UNet	Blast-Net Rad et al. (2019a)	0.0024	True
SP-Swin UNet	Trans UNet Chen et al. (2021)	0.4952	False
SP-Swin UNet	Swin UNet Cao et al. (2022)	0.6081	False

model performance and the ecological aspects of the experimentation are presented. The *Dice Score or F*1 values for fire segmentation indicate the accuracy of the models in delineating fire regions. Both the mean and standard deviation of Dice Score or F1 are reported, providing insights into the consistency and reliability of the segmentation results. Additionally, the table includes information on the Energy Consumption (EC) and carbon footprint (CO2e) of the models. This transparency is important to promote sustainable machine learning applications by considering the environmental impact of training and running the models. Quantitative values for *Recall* providing information about FN are reported in Table 1 which are of particular importance in the context of fire.

Furthermore, in our study of the SP-Swin UNet model, we investigated the individual contributions of its two key components: CAPP and C^3. To highlight the significance of these components, we conducted two separate experiments, eliminating one component at a time.

4.2.2. Statistical Significance

To validate the significance of these differences, we employed statistical tests, specifically an analysis of variance (ANOVA) followed by Tukey's Honestly Significant Difference (HSD) test. The post hoc test results provide insights into which pairs of methods have statistically significant performance differences and which differences might be considered marginal. While certain performance differences may not exhibit statistical significance, it's important to note that even modest improvements hold the potential to yield life-saving outcomes in the critical context of wildfire detection.

4.2.3. Performance and Training Size

Figure 7 depicts the performance of various models in relation to the training ratio. The training ratio refers to the proportion of available data used for training the models. This figure highlights the models' ability to learn from limited data. Even with very little training data, some models exhibit better learning capabilities compared to others, indicating their potential for efficient knowledge extraction and utilization. The proposed *SP-Swin UNet* achieves superior performance for all training ratios considered.

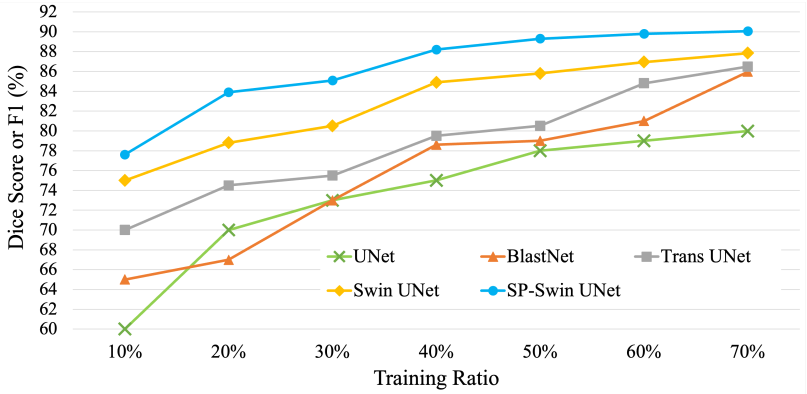


Figure 7: Performance comparison of models with different training ratios.

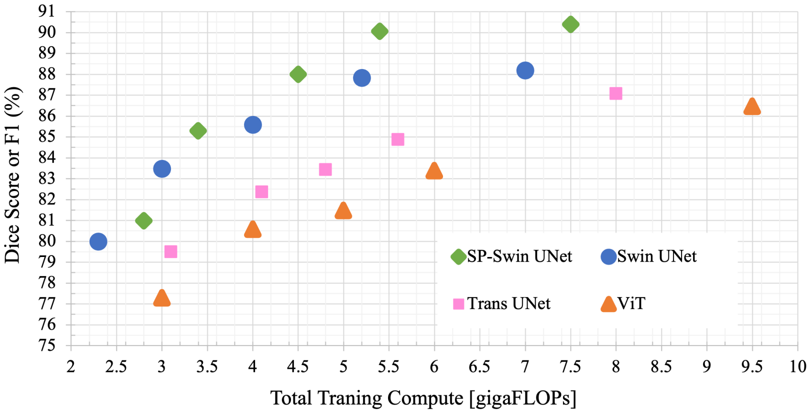


Figure 8: Performance versus training compute for different models: Vision Transformers Dosovitskiy et al. (2020), Swin UNet Cao et al. (2022), Trans UNet Chen et al. (2021), and our SP-Swin UNet.

4.2.4. Performance Gains and Computational Demands

Our experimental results demonstrate the effectiveness of the *SP-Swin UNet* vision transformer in improving wildfire detection accuracy, particularly in scenarios where long-range dependencies are essential. However, it is important to consider the potential trade-offs introduced by the increased computational complexity compared to other models considered. Figure 8 contains the performance versus total training compute. First, it can be observed that *SP-Swin UNet* outperforms *ViT* and *Trans UNet* on the performance/compute tradeoff. Second, *SP-Swin UNet* uses approximately $10\% - 15\%$ more compute to attain the same performance as *Swin UNet*. Third, at larger computational budgets *SP-Swin UNet* outperforms *Swin UNet* by 2% in Dice Score, but that improvement comes with 8% more computational demand.

4.2.5. Latency and Throughput

Furthermore, the throughput in terms of images processed per second (image/s) is measured. This metric provides insights into the model's efficiency during inference which is particularly relevant to real-time applications such as wildfire detection. It turns out that our *SP-Swin UNet* vision transformer achieves an inference throughput of 681.8 images/sec, surpassing the performance of the *ViT* model, which achieves 632.1 images/sec by a margin of 50. Furthermore, the *SP-Swin UNet* falls short of *Swin UNet* (with 715.3 images/sec.) by a mere 35 images/sec.

4.3. Qualitative Results

Figure 9 presents the qualitative results comparing the fire identification maps generated by the proposed *SP-Swin UNet* model and those of *Trans UNet* Chen et al. (2021) and *Swin UNet* Cao et al. (2022). The figure consists of seven rows, each representing a sample patch. The first column (column a) displays false-color Landsat-8 patches and the second column (column b) shows the green channel of the input images providing visual context for the subsequent columns. Columns c and d showcase the results of *Trans UNet* Chen et al. (2021) and *Swin UNet* Cao et al. (2022), respectively, representing alternative segmentation approaches for comparison. Finally, column e displays the fire prediction map generated by the SP-Swin UNet model, indicating the areas identified as fire. The qualitative results provide a visual assessment of the model's performance in accurately delineating fire regions and highlight the effectiveness of the proposed approach in fire segmentation tasks.

In the context of our analysis, the different categories of pixel classifications are visually represented using distinct colors. True negatives (TN) are represented by the color black, indicating correct identification of non-fire regions. True positives (TP) are visualized in white, representing accurate detection of the fire regions. False negatives (FN) are depicted in blue, indicating instances where the fire regions were present but not detected. Lastly, false positives (FP) are shown in red, representing cases where the fire regions were incorrectly identified despite their absence. It is important to note that the impact of a single false negative in fire mask segmentation cannot be underestimated. Even a small fire, if undetected, can quickly spread and grow, leading to catastrophic outcomes.

5. Conclusion

In conclusion, this research paper proposed a robust approach for detecting active wildfires using multispectral satellite imagery in North America. While we focused on North America due to the urgency of addressing the alarming increase in wildfires and their impacts, we recognize the importance of adapting and assessing our approach for different parts of the world with varying wildfire characteristics. By leveraging vision transformers and a

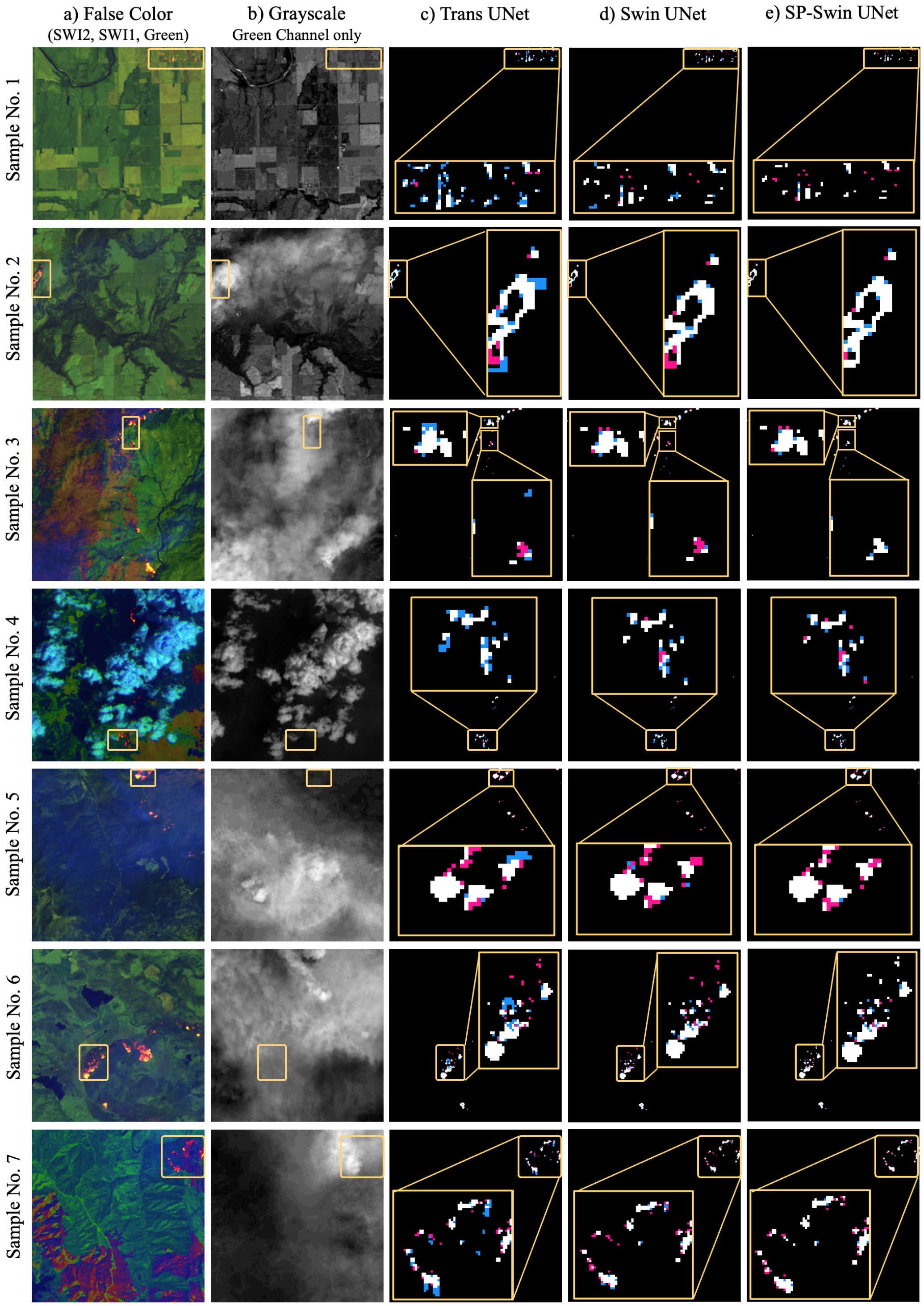


Figure 9: Qualitative comparison of the Blast-Net and the state-of-the-art models. Here, white represents True positives (TP), blue represents False negatives (FN), and red represents False negatives (FN).

vast repository of landsat-8 satellite data, the study aimed to address the increasing size and severity of wildfires in the region. A novel u-shape vision transformer was proposed to effectively capture spatial dependencies and learn meaningful representations from multispectral images. Vision transformers have emerged as a promising alternative to CNNs, leveraging the power of self-attention mechanisms to capture global dependencies and long-range interactions in images. Vision transformers have demonstrated promising results in learning from limited amounts of data. However, training Vision transformers can become challenging and it often requires pre-training on large-scale datasets, such as ImageNet, to achieve their full potential.

References

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 205–218, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-25066-8.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, abs/2102.04306, 2021. URL https://arxiv.org/abs/2102.04306.

L. C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Gabriel Henrique de Almeida Pereira, Andre Minoro Fusioka, Bogdan Tomoyuki Nassu, and Rodrigo Minetto. Active fire detection in landsat-8 imagery: A large-scale dataset and a deep-learning study. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:171–186, 2021. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2021.06.002. URL https://www.sciencedirect.com/science/article/pii/S092427162100160X.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. IEEE CVPR*, pages 7132–7141, 2018.

V. Iglovikov and A. Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, June 2022.

W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Info. Process. Syst.*, pages 4898–4906, 2016.

R. M. Rad, P. Saeedi, J. Au, and J. Havelock. Multi-resolutional ensemble of stacked dilated u-net for inner cell mass segmentation in human embryonic images. In *IEEE Int. Conf. on Image Process. (ICIP)*, pages 3518–3522. IEEE, 2018a.

R. M. Rad, P. Saeedi, J. Au, and J. Havelock. Blastomere cell counting and centroid localization in microscopic images of human embryo. In *Proc. IEEE Int. Workshop on Multimedia Signal Process.*, pages 1–5, 2018b.

R. M. Rad, P. Saeedi, J. Au, and J. Havelock. Multi-resolutional ensemble of stacked dilated u-net for inner cell mass segmentation in human embryonic images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3518–3522, 2018c. doi: 10.1109/ICIP.2018.8451750.

R. M. Rad, P. Saeedi, J. Au, and J. Havelock. Blast-net: Semantic segmentation of human blastocyst components via cascaded atrous pyramid and dense progressive upsampling. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1865–1869, 2019a. doi: 10.1109/ICIP.2019.8803139.

R. M. Rad, P. Saeedi, J. Au, and J. Havelock. Cell-net: Embryonic cell counting and centroid localization via residual incremental atrous pyramid and progressive upsampling convolution. *IEEE Access*, 7:81945–81955, 2019b. doi: 10.1109/ACCESS.2019.2920933.

O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, pages 234–241. Springer International Publishing, Cham, 2015. ISBN 978-3-319-24574-4.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020. URL https://arxiv.org/abs/2012.12877.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. *arXiv preprint arXiv:1702.08502*, 2017.

F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *Proc. Int. Conf. on Learning Representations*, 2016.

F. Yu, V. Koltun, and T. A. Funkhouser. Dilated residual networks. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognition*, volume 2, page 3, 2017.

Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. *CoRR*, abs/2102.08005, 2021. URL https://arxiv.org/abs/2102.08005.

H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognition*, pages 2881–2890, 2017.