SEMANTIC SEGMENTATION ON THE LANDSLIDE4SENSE DATASET

Reiko Lettmoden

Data Science
Technische Universität Braunschweig

Amit Amit

Data Science
Technische Universität Braunschweig

ABSTRACT

In 2022, the Institute of Advanced Research in Artificial Intelligence organized the Landslide4Sense (L4S) competition, focusing on landslide detection via large-scale satellite imagery. This study harnesses semantic segmentation and deep learning models to automatically identify landslides within the L4S dataset, encompassing diverse global regions affected by landslides from 2015 to 2021. Through extensive experimentation, we explore various configurations, including band selection, augmentations, loss functions, and model architectures. Key findings reveal the effectiveness of Lovazs loss, the impact of augmentations on performance improvement, and the critical role of band selection in UNet and UNet-Swin models. These insights provide a foundation for advancing landslide detection techniques, contributing to disaster management and remote sensing applications.

Index Terms— Semantic Segmentation, Landslide Detection, Remote Sensing, Deep Learning, Landslide4Sense Dataset

1. INTRODUCTION

Landslides, characterized by the downward movement of rock, debris, soil, or a combination thereof, represent a significant geological hazard responsible for extensive damage to infrastructure and the potential loss of lives. To mitigate these devastating impacts, the accurate and automated detection of landslide-prone areas is imperative. In response to this challenge, we delve into the realm of deep learning and semantic segmentation to address this critical issue.

The Landslide4Sense (L4S) dataset, sourced from diverse regions worldwide and spanning the years 2015 to 2021, forms the foundation of our investigation. This rich dataset, comprising multispectral data from Sentinel-2, slope data from ALOS PALSAR, and digital elevation models (DEMs), is the cornerstone of our efforts to automate landslide detection [1].

Motivated by the pressing need for precise identification and mapping of landslide-affected regions, we embark on a comprehensive journey. Our exploration encompasses various aspects, including band selection, augmentations, loss functions, and model selection. We aim to unravel the most effective strategies and configurations for leveraging semantic segmentation in this critical domain.

Our findings offer insights into the utilization of satellite imagery for landslide detection, highlighting the significance of the Lovazs loss, the benefits of augmentations, and the impact of band selection on model performance. Ultimately, our research not only contributes to the advancement of landslide detection methodologies but also underscores the potential of semantic segmentation in remote sensing applications, with implications for disaster management and beyond.

2. DATASET

The Landslide4Sense (L4S) dataset is a pivotal resource for advancing the field of landslide detection through semantic segmentation using remote sensing data. Its key features include:

2.1. Data Splits

The L4S dataset is divided into three primary splits: training, validation, and test, with a respective distribution of 60%, 20%, and 20% of the available 3799 image patches. We construct our dataset by choosing half of the train split and split the obtained data in the same ratios into training, validation and test. Each image patch measures 128 x 128 pixels and is labeled pixel-wise to indicate landslide-affected areas.

2.2. Data Bands

The dataset encompasses a rich collection of 14 bands, providing comprehensive information for landslide detection. These bands include [1]:

Multispectral Data from Sentinel-2 (B0 to B11): This multispectral data captures a wide range of spectral information, enabling precise discrimination of different land cover classes and contributing to the accuracy of landslide detection.

Slope Data from ALOS PALSAR (B12): Incorporating slope data is essential for understanding the topographic characteristics of the landscape, a crucial factor in identifying potential landslide-prone areas.

Digital Elevation Model (DEM) from ALOS PALSAR (B13): The DEM data enriches the dataset by providing detailed topographical information, aiding in the accurate mapping and detection of landslides.

2.3. Dataset EDA

All bands in the L4S dataset have been resized to a consistent resolution of approximately 10 meters per pixel. This standardized resolution ensures uniformity and compatibility across the dataset.

Addressing Data Imbalance: One of the notable challenges in landslide detection is class imbalance. Landslides are relatively rare compared to non-landslide areas, leading to an imbalance in the dataset as shown in the Figure 1. To mitigate this issue, several strategies have been explored, including the use of weighted loss functions, such as the weighted cross-entropy (WCE) loss and Lovazs loss. These techniques aim to give more importance to the minority class (landslides) during model training, improving the model's ability to detect landslides accurately.

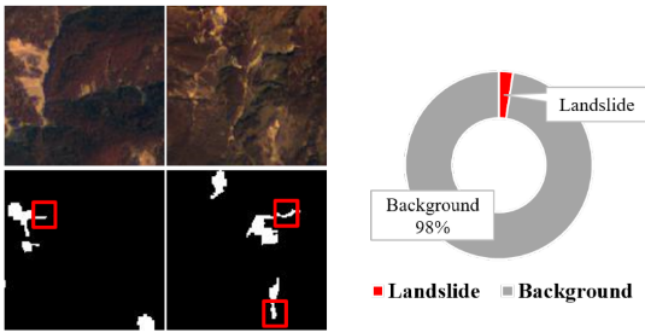


Fig. 1. Small landslide branches seen in the red boxes (a) and class imbalance (b) by Ghorbanzadeh et al. [1].

2.4. Metrics

The evaluation process hinges on a set of essential metrics that offer insights into the model's effectiveness. These metrics include precision(eq. 1), recall(eq. 2), F1-Score(eq. 3), and Intersection over Union (IoU)(eq. 4). Precision quantifies the model's ability to minimize false positives, while recall measures its capacity to correctly identify actual landslide areas. The F1-Score strikes a balance between precision and recall, offering a comprehensive assessment of the model's performance. IoU evaluates the model's precision in delineating precise landslide boundaries, ensuring accurate segmentation.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3}$$

$$IoU(U, V) = Jaccard(U, V) = \frac{|U \cap V|}{|U \cup V|} \tag{4}$$

3. METHODS

As baseline for our work, we use a UNet[2] which is also used as baseline in the challenge [3]. Our aim is to cover methods from the lecture such as model selection, losses and augmentations to improve the baseline.

As a motivation to investigate different input bands, figure 2 shows the influence of band selection on different architectures. The UNet[2] performes best on all input bands and beats both the more recent Deeplabv3[4] and Swin Transformer[5]. While the Deeplabv3 performs worse for all band selections, the Swin Transformer performs best of all models on RGB bands only while it performs significantly worse than the UNet on all bands. Since we use another dataset, we perform experiments with different input bands on the UNet and later on the Swin Transformer.

		F1 Score (%)		
Input spectral bands	Input bands	Swin Transformer	Deeplabv3	U-Net
RGB	3	**65.6**	58.0	59.2
SWIR	3	55.6	50.2	52.1
NGB	3	60.8	**59.2**	58.9
PCA [44]	3	49.5	46.8	52.4
RGB + NIR	4	63.3	57.2	59.4
RGB + SWIR	6	58.2	55.9	59.8
RGB + NIR + SWIR	7	54.8	57.5	60.0
All bands	14	55.8	57.8	**61.1**

Note: In this table, the RGB denotes the red, green, and blue spectra. SWIR denotes the 3-band far infrared in Sentinel-2. NGB denotes the near-infrared, green, and blue spectra. NIR denotes the near-infrared spectral. PCA refers to the techniques [44] of dimensionality reduction for compressing the original 14 bands into 3 bands.

Fig. 2. Influence of selected input bands on the F1-Score of the Swin Transformer, Deeplabv3 and UNet by Ghorbanzadeh et al. [1].

Augmentations are known to avoid models to overfit less on training data and are used to generate data with barely any costs. We apply common augmentations such as ColorJitter, Rotation by 90 degrees and both vertical and random flip. For the ColorJitter augmentation we follow competitors of the challenge and apply it only to multispectral bands and not to any height bands [1].

As seen in figure 1, our data is highly imbalanced with the minority class making up 2% of the data. Amongst other methods such as resampling, losses are used to avoid highly

biased models towards the majority class. The cross entropy (CE) loss is used as default loss.

The weighted cross entropy (WCE) loss assigns each class a weight, such that minority classes can be upweighted. Equation 5 shows the formula for the WCE loss, where p denotes the prediction probability vector of an sample and y the label.

$$WCE(p, y) = -\sum_{c=1}^{M} y_c \log(p_c) w_c \tag{5}$$

M are all the classes which are iterated over and w_c the weight for a class. By setting all the weights to 1, the CE loss is obtained.

The Lovasz loss [6] is used by many competitors of the landslide challenge[1] which is why we also apply the Lovasz loss. It directly optimizes the IoU indices and thus mIoU which is a more pessimistic metric than the F1-Score.

The first two places of the challenge apply a Swin Transformer to utilize a more recent and more powerful architecture[1]. Figure 3 shows on the top the combination of the UNet with a Swin Transformer, where the encoder blocks of the UNet are replaced with the blocks of the Swin Transformer. The Swin Transformer is a strong vision enconder which is applied to many computer vision tasks[5].

Additionally, the first place uses a SegFormer[7] with a different architecture than the UNet, as seen in the lower illustration of figure 3. Instead of an U-shape, each encoder block is forwarded through a MLP layer and then concantenated in the decoder. The decoder is made of 4 MLP layers and thus allows for light-weight architectures, such as the SegFormer-B0 with 4.1 million parameters only.

4. IMPLEMENTATION

GeoSeg [1] provides a training framework based on *PyTorch* and *PyTorch Lightning*. The SegFormer and Swin Transformer architectures are implemented by *timm* and an UNet implemented in PyTorch[2].

AdamW[8] is used as optimizer with learning rate set to 6e-5 and weight decay to 0.01. For loss scheduling, CosineAnnealingLR is applied with a minimum learning rate set to 1e-6 for all models except the SegFormer model. The minimum learning rate for the SegFormer model is set to 1e-7.

We employ early stopping with a patience of 15 epochs and monitor the F1-Score of the validation set. The batch size is set to 16 and we train for a maximum of 100 epochs. All training are done on a T4 GPU with Google Colab.

In our experiments, we use the SegFormer-B0 and Swin Transformer tiny (SwinT). For the SegFormer-B0 we resize the input image to 512x512 as suggested by the first place[1]. However, training time gets too long with large input sizes for the larger SwinT UNet such that we only resize the image to 256x256. We upscale the labels with nearest neighbour interpolation while we upscale the images with bilinear interpolation.

[1]https://github.com/WangLibo1995/GeoSeg

[2]https://github.com/milesial/Pytorch-UNet

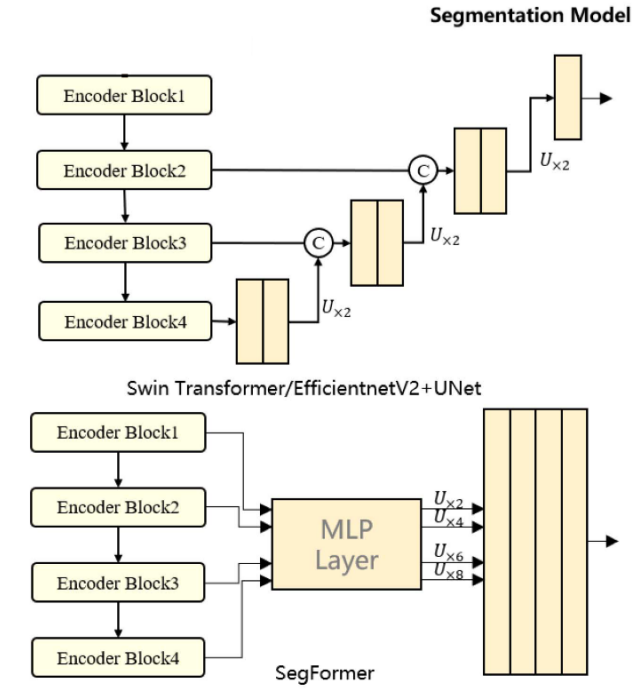


Fig. 3. Architecture illustration of the Swin Transformer based UNet (top) and SegFormer (bottom) by Ghorbanzadeh et al. [1].

As initial baseline we use a UNet with transposed convolutions, cross entropy loss and no augmentations. In each subsection of chapter 5, we experiment with different configurations for a method and set the best performing configuration as new baseline for the upcoming ablation studies.

5. RESULTS

5.1. Band Selection

Table 1 shows the F1-Scores for different subsets of bands on our test split. The baseline using all bands achieves the best results and the model trained on RGB faces a highest decline in F1-score. On the other hand, adding height information with bands 12 and 13, the model achieves almost the same F1-Score as the UNet trained on all bands. Adding another near infrared band does not significantly change the F1-Score. Hence, the restricting ourselves to RGB and Height bands for the UNet would already simplify the amount of data required without dealing with a large performance drop.

In section 5.4, we take a look at the Swin-UNet trained on all bands and RGB bands only, as the Swin-UNet performs best on RGB bands only (compare figure 5.1).

Name	Bands	F1-Score
Base	ALL	**0.721**
RGB	2, 3, 4	0.676
RGB Height	2, 3, 4, 12, 13	0.718
RGB Height NIR	2, 3, 4, 5, 12, 13	0.715

Table 1. F1-Scores (Test) of an UNet trained on different input bands.

5.2. Augmentations

Figure 4 shows, that the UNet without augmentations (orange) starts overfitting after roughly 20 epochs. The train F1-Score continues to increase while the validation F1-Score decreases, such that training is stopped due to early stopping. The UNet with augmentations avoids overfitting, as both validation and train F1-score are increasing. The test F1-Score increases by 0.018 for the UNet with augmentations such that augmentations are used from now on as a baseline.

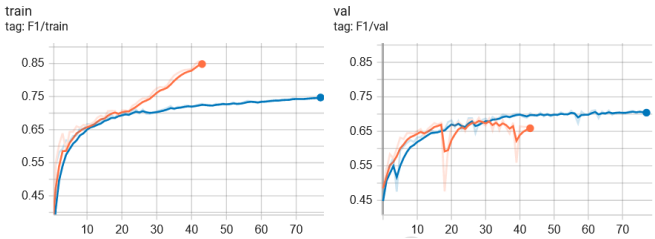


Fig. 4. F1-Score on train and validation set of an UNet trained without augmentations (orange) and with augmentations (blue).

5.3. Losses

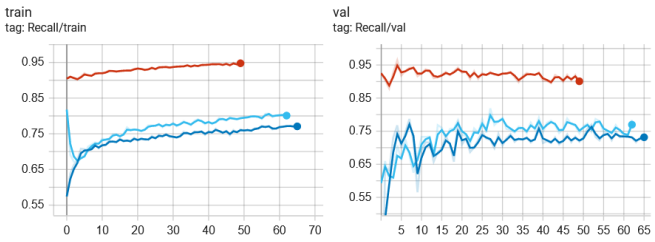


Fig. 5. Recall on train and validation set of an UNet trained with a Lovasz loss (dark blue), weighted cross entropy (WCE) loss with weight 10 for the minority class (brown) and with WCE weight 3 for the minority class (bright blue).

For the WCE loss, the weight for the majority class (non-landslide) is set to 1 whereas the weight for the minority class is set to [2, 3, 5, 10] in different experiments. Higher weights for the minority classes increase the recall, as seen in table 2 and figure 5. Hence, choosing a weight of 10 for the minority class, increases recall in comparison to the CE baseline from 0.67 to 0.89.

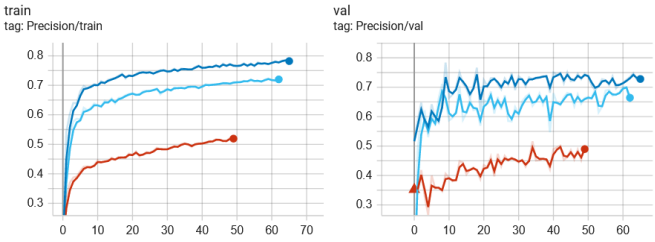
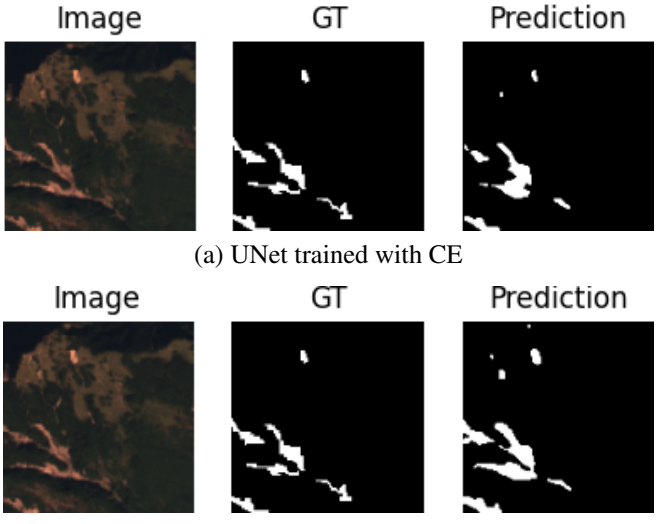


Fig. 6. Precision on train and validation set of an UNet trained with a Lovasz loss (dark blue), weighted cross entropy (WCE) loss with weight 10 for the minority class (brown) and with WCE weight 3 for the minority class (bright blue).

As a trade-off, the precision decreases as the model predicts more false positives. This can be seen in figure 6 and table 2 where the *WCE 1-10* model achieves the highest recall and the lowest precision. An qualitative example is given in figure 7 in which the UNet trained with WCE classifies more pixels as landslide in each cluster. Thus, the *WCE 1-10* model has with 0.55 significantly less precision than the unbiased model with 0.81 precision. On the other hand, the Lovasz loss increases recall by 0.1 while having 0.04 less precision than the CE model. Hence, it achieves the best test F1-Score of all models such that from this point on the Lovasz loss is employed as standard loss.



(a) UNet trained with CE

(b) UNet trained with WCE 1-10

Fig. 7. Qualitative example of an UNet trained with different losses.

Loss	Recall	Precision	F1-Score
CE	0.673	**0.811**	0.736
WCE 1-2	0.776	0.716	0.745
WCE 1-3	0.782	0.715	0.747
WCE 1-5	0.839	0.628	0.718
WCE 1-10	**0.893**	0.553	0.683
Lovasz	0.741	0.771	**0.756**

Table 2. Recall, Precision and F1-Scores (Test) of an UNet trained with different losses.

5.4. Architecture

Figure 8 shows the F1-Score for different architectures. The SegFormerB0 converges quickly, such that the other architectures are trained longer. Table 3 shows that the SegFormer performs worse than the UNet but still achieves with roughly 13% the parameters of an UNet a good performance. The SwinT UNet trained on RGB bands overftis the most since it achieves the highest train F1-Score but the lowest test F1-Score. Here, more data might be needed to properly train the architecture. The SwinT UNet trained on all bands still overfits but performs significantly better on the test and validation F1-Score.

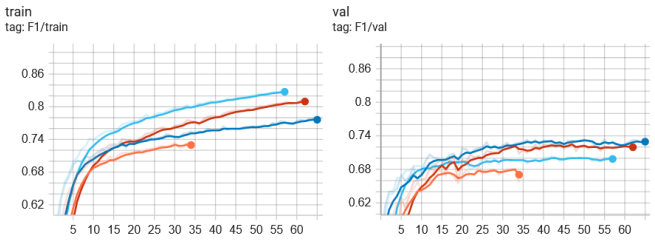


Fig. 8. F1-Score on train and validation set of an SegFormerB0 (orange), UNet (dark blue), SwinT UNet (brown) and SwinT UNet trained on RGB bands only (bright blue).

Architecture	Resize	Bands	F1-Score
UNet	-	ALL	**0.756**
SwinT UNet	256x256	ALL	0.745
SwinT UNet	256x256	RGB	0.726
SegFormerB0	512x512	ALL	0.728

Table 3. F1-Scores (Test) of different architectures.

5.5. Overview

An summary for all categories of ablation studies is shown in table 4. All experiments are shown in table 5.

Bands	Augs	Loss	Architecture	F1-Score
ALL	-	CE	UNet	0.721
RGB Height	-	CE	UNet	0.718
ALL	+	CE	UNet	0.736
ALL	+	Lovasz	UNet	**0.756**
ALL	+	Lovasz	SwinT UNet	0.745

Table 4. Summary of ablation studies with F1-Scores (Test).

6. LIMITATIONS & FUTURE WORK

Our work is restricted by the computation power due to the free Google Colab. Ghorbanzadeh et al. [1] propose to resize the data to 512x512 images, which becomes computationally infeasible for training. Additionally, we are restricted to the smallest SegFormer model which has significantly less parameters than the UNet and makes the comparison of architectures more difficult.

The computation limitations result in the choice of a subsplit in the dataset. Since we use half of the training data from the landslide4sense train split, we observe overfitting in the Swin Transformer based UNet, especially for the configuration with RGB input bands only. These models might benefit from more data such that the whole dataset can be investigated future work.

Since we restrict ourselves to the train data of the challenge, test and validation data of the challenge are not utilized to improve our models. The test and validation splits of the challenge are taken from different geographic places than the train set. Hence, our models have not to deal with larger distribution shifts and do not suffer from major declines in test F1-Scores. It is thus difficult to compare our work to the challenge which can be done in future work by utilizing the whole dataset.

7. CONCLUSION

In this work, we train models for semantic segmentation on a subset of the L4S train split to identify landslides.

We improve the UNet baseline by applying augmentations to avoid overfitting. By choosing different losses, the model is able to learn from the imbalanced dataset and thus achieve the highest increase of F1-Score. Additionally, we show the influence on the precision-recall trade-off based on different losses.

Future work can utilize all available L4S data to better compare more powerful architectures with the UNet and to train models with better generalizability.

8. REFERENCES

[1] Omid Ghorbanzadeh, Yonghao Xu, Hengwei Zhao, Junjue Wang, Yanfei Zhong, Dong Zhao, Qi Zang, Shuang

Name	Bands	Loss	Augs	Model	Resize	F1-Score
ALL	ALL	CE	-	UNet	-	0.721
RGB	2, 3, 4	CE	-	UNet	-	0.676
RGB Height	2, 3, 4, 12, 13	CE	-	UNet	-	0.718
RGB Height NIR	2, 3, 4, 5, 12, 13	CE	-	UNet	-	0.715
ALL Aug	ALL	CE	+	UNet	-	0.736
Lovazs	ALL	Lovazs	+	UNet	-	**0.756**
Lovazs 256	ALL	Lovazs	+	UNet	256	0.755
WCE 2	ALL	WCE 1-2	+	UNet	-	0.745
WCE 3	ALL	WCE 1-3	+	UNet	-	0.747
WCE 5	ALL	WCE 1-5	+	UNet	-	0.718
WCE 10	ALL	WCE 1-10	+	UNet	-	0.683
SwinT UNet RGB 256	2, 3, 4	Lovazs	+	SwinT UNet	256	0.726
SwinT UNet 256	ALL	Lovazs	+	SwinT UNet	256	0.745
SegFormerB0 512	ALL	Lovazs	+	SegFormerB0	512	0.728

Table 5. Overview of all configurations with F1-Score on test set.

Wang, Fahong Zhang, Yilei Shi, et al., "The outcome of the 2022 landslide4sense competition: Advanced landslide detection from multisource satellite imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9927–9942, 2022.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[3] Omid Ghorbanzadeh, Yonghao Xu, Pedram Ghamisi, Michael Kopp, and David Kreil, "Landslide4sense: Reference benchmark data and deep learning models for landslide detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[6] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4413–4421.

[7] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021.

[8] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.