Comparative Analysis of Pre-trained models across diverse data Modalities

Team Name
512Project 1

Amogh Agashe
115634589
aagashe@cs.stonybrook.edu

Preet P. Amin
115360499
pamin@cs.stonybrook.edu

Rishabh V. Patil
115808342
rishabhvijay.patil@stonybrook.edu

Ameya U. Zope
115940930
ameya.zope@stonybrook.edu

Abstract

This project investigates biases in various deep learning pre-trained models including DeepLabV3+, SETR and SegFormer designed for semantic segmentation tasks in computer vision. Semantic segmentation involves assigning a class label to each pixel in an image, thus segmenting the image into different regions based on the content or objects present. We evaluate these models on a range of different data types, including RGB, grayscale, and inverted grayscale images. Our main objective is to comprehensively analyze the factors influencing the performance of these models when applied to diverse datasets. By doing so, we aim to extend the applicability of these pre-trained models to different data modalities beyond those on which they were originally trained. The insights derived from our results offer a clear understanding of the capabilities and weaknesses of various pre-trained models, providing valuable information for researchers and practitioners in the field of Machine Learning.

Introduction

Computer Vision is the field of artificial intelligence and computer science that focuses on enabling machines to understand and interpret visual information. Deep learning has been widely applied to solve many vision tasks such as Image Classification, Object Detection, Image Segmentation, Image Transformation, etc. Image Segmentation is a fundamental Computer Vision task in which we partition the image into meaningful parts such that each segment corresponds to an object or area of interest. Many downstream tasks of Image Segmentation are connected, so it is natural to assume that a model pre-trained on one dataset can be useful for another.

It is considered a general practice to use pre-training and transfer learning to finetune an existing model for a specific downstream task[10]. For example, the backbone of Image Segmentation tasks[5,6,1-3] is often pre-trained on the ImageNet dataset[11]. A lot of such models are easily available such as ResNet[8], U-Net[5], DeepLabV3+[6], etc. In light of the recent success of Transformers in various tasks of NLP,

Dosovitskiy et al. [11] introduced Vision Transformer which treats a 16x16 image patch as a word. Since then there have been various attempts to efficiently use Transformers for Visual tasks[1,3]. Spatial-Enhanced Transformer (SETR) [12] introduces spatial embeddings to capture positional information, providing a robust mechanism for handling spatial relationships in semantic segmentation tasks. Its architecture combines global and local context information, enhancing the model's ability to make precise pixel-wise predictions across diverse datasets. SegFormer[1] represents another significant advancement in the fusion of transformers with computer vision which uses Transformer Pyramid to extract features from images at multiple resolutions. This allows SegFormer to use a lightweight decoder and still maintain high performance.

The aforementioned pre-trained models are largely trained on RGB images (which are also captured by sampling RGB waves from the real world). The straightforward applications of downstream tasks directly use the features produced by these pre-trained modes and only train the decoder or classification layers. However, this practice inherently exposes the models to biases present in the data distribution of RGB waves from which they were originally trained. Consequently, questions arise regarding the performance of these pre-trained models when applied to different input distributions, such as Grayscale and inverted Grayscale. Examining these two modalities is crucial as it provides insights into how these models operate in the absence of color information, particularly in scenarios like dark/night images, infrared images, and even in the context of X-rays and inverted X-rays. This exploration aims to shed light on the adaptability and performance of the models under diverse circumstances and input variations.

In our study, we compare three different pre-trained models including DeepLabV3+, SETR and SegFormer, and execute image segmentation tasks on cityscape dataset images. This dataset contains city images with different classes including buildings, people, cars, trucks, road, sky, etc. Our objective is to systematically evaluate the performance of the stated above pre-trained models on non-conventional image distributions i.e grayscale and inverted grayscale. This comprehensive study aims to provide insights into the efficiency of these models for previously unseen data modalities and provide a reasonable explanation on any significant discrepancies seen during performing the experiments and observing the results.

Through the conducted experiments we achieve the following conclusions:

- Understanding the key differences between the functionalities of different models on different data modalities.
- Investigate why Segformer works better on Inverted Grayscale images in comparison to SETR and DeepLabV3+.
- Determine why re-training is preferable to fine-tuning when the data distribution exhibits significant variability.

Related Work

DeepLabv3+ [6] is an advanced semantic segmentation architecture, building upon DeepLabv3 [7], with notable enhancements such as an effective decoder module for refining segmentation results. Developed at Google, the DeepLab series has evolved through versions like DeepLabv1 [17], DeepLabv2 [16], and DeepLabv3 [7]. It utilizes Atrous Spatial Pyramid Pooling (ASPP) and an Encoder-Decoder Architecture to encode multi-scale contextual information and recover location and spatial data. The network features the Modified Aligned Xception and Atrous Separable Convolution for improved speed and robustness. Widely acclaimed for high-quality image segmentation, DeepLabV3+ finds applications in object detection, image labeling, and scene parsing, particularly in tasks requiring fine-grained object delineation and precise

localization [9]. The architecture employs ResNet 101 for encoding. ResNet-101, a Residual Network member [8], stands out for its depth with 101 layers, addressing the vanishing gradient problem through residual blocks and skip connections. It was pre-trained end-to-end using the Imagenet dataset. Utilizing a bottleneck architecture within each residual block, it optimizes computational efficiency while enhancing representational capabilities. ResNet-101 is widely adopted in computer vision tasks like image classification, object detection, and semantic segmentation due to its effectiveness and depth.

SETR (Segmentation Transformer) [12] offers a novel approach to semantic segmentation, addressing some of the limitations of traditional models like Fully Convolutional Networks (FCNs). Traditional FCN-based models work by processing images in a step-by-step manner, gradually reducing the image's resolution. This is done to capture more abstract information but limits their ability to understand long-range relationships in complex scenes, which are crucial for semantic segmentation in real-world images.

SETR introduces a new way of looking at semantic segmentation. Instead of using a series of convolutional layers for image understanding, it employs a Transformer, a type of model known for its success in natural language processing. In SETR pre-training, the model initializes transformer layers and input linear projection using pre-trained weights from ViT or DeiT, while FCN encoder employs ImageNet-1k pre-training. Patch size 16×16 is used, and 2D interpolation adjusts pre-trained position embeddings during fine-tuning for various input sizes. SETR breaks the input image into small patches and transforms them using a Transformer. Each patch is represented by a feature vector, and these vectors are used as input for the Transformer. Unlike traditional models, this model doesn't reduce the image's resolution during the encoding process. It maintains the full spatial information while employing global self-attention mechanisms. In practical tests, SETR has proven its effectiveness. It not only offers a new way of designing models but also achieves state-of-the-art results on benchmark datasets for semantic segmentation.

SegFormer[1] uses the previously unconventional approach of using Transformers as an encoder for the image segmentation task. They pre-train the encoder on the Imagenet-1K dataset and randomly initialize the decoder. SegFormer takes advantage of the Transformer model's self-attention mechanism, which has proven to work well on NLP tasks, to capture long-distance dependencies on Images. SegFormer is structured around two primary components: (1) A hierarchical Transformer encoder that generates high-resolution coarse features and low-resolution fine features. (2) A lightweight All-MLP decoder that merges these multi-level features to generate the semantic segmentation mask. Contrary to ViT[2], SegFormer uses a pyramid structure in a Transformer similar to PVT[3] to capture multi-level features at different image resolutions. To reduce the complexity of calculations, self-attention is only calculated in a restricted neighborhood around the input patch. To overcome the challenges introduced by the PE (Positional Encoding) scheme, SegFormer replaced the PE by taking advantage of the inherent property of Convolution layers. In SegFormer, the Mix-FNN component takes into account the impact of zero padding to retain essential location information [4]. It achieves this by incorporating a 3×3 convolution in the feed-forward network (FFN). Due to the Transformer Pyramid's larger effective receptive field, SegFormer can employ a much simpler Decoder scheme of using all-MLP Decoder.

"Pre-training on Grayscale ImageNet Improves Medical Image Classification"[13] explores the challenges in using pre-trained models (trained on mostly RGB images, for example ImageNet), for medical image analysis due to the inherent color discrepancy between natural and medical images. Apparently, fine-tuning RGB pre-trained models on Grayscale images carry artifacts and biases from their pre-training and affect the overall performance. In this paper, they convert ImageNet images to grayscale for pre-training and surprisingly, this grayscale pre-trained model not only maintains performance on the original ImageNet classification task but also outperforms color ImageNet models when fine-tuned for disease classification in

chest X-ray images. The research highlights the benefits of this strategy, demonstrating improved speed, accuracy, and elimination of unnecessary pre-processing steps, suggesting that grayscale pre-training allows the model to learn features more pertinent to medical datasets.

“Impact of Colour on Robustness of Deep Neural Networks” [14] emphasizes that the robustness of deep neural networks in image classification tasks when trained on RGB image datasets and exposed to color distortions is a growing concern in computer vision research. This paper delves into this issue, shedding light on the dramatic reduction in model performance when deep convolutional neural networks are presented with color-distorted images. Notably, this study reveals transforming color distributions in input image (such as deleting a RGB channel, drastically changing hue and saturation etc.) significantly impairs the performance of pre-trained models.

“A comparative study of X-ray and CT images in COVID-19 detection using image processing and deep learning techniques” [18] highlights the difference between X-ray images, CT scans Images and RGB images. X-ray and CT imaging are the two often utilized medical imaging modalities for detecting COVID-19. But due to differences in color distribution and data artifacts, the borders and edges of the images are not clear which may lead to a false diagnosis of the disease. The authors explore different pre-processing techniques (such as histogram enhancement) and/or augmentation techniques to overcome data scarcity problems with pre-trained models like VGG and ResNet 50.

Methodology

The methodology of this project seeks to systematically evaluate and compare the performance of three prevalent pre-trained models—DeepLab V3+, SETR, and SegFormer, across distinct data modalities. Firstly, the reason behind comparing pre-trained models lies in addressing potential biases ingrained during their initial training on large-scale RGB datasets. These biases might influence their proficiency when dealing with diverse data types, hence prompting the necessity for a comparative analysis. The chosen models represent different architectures and attention mechanisms, facilitating a comprehensive assessment of their adaptability to varied data distributions.

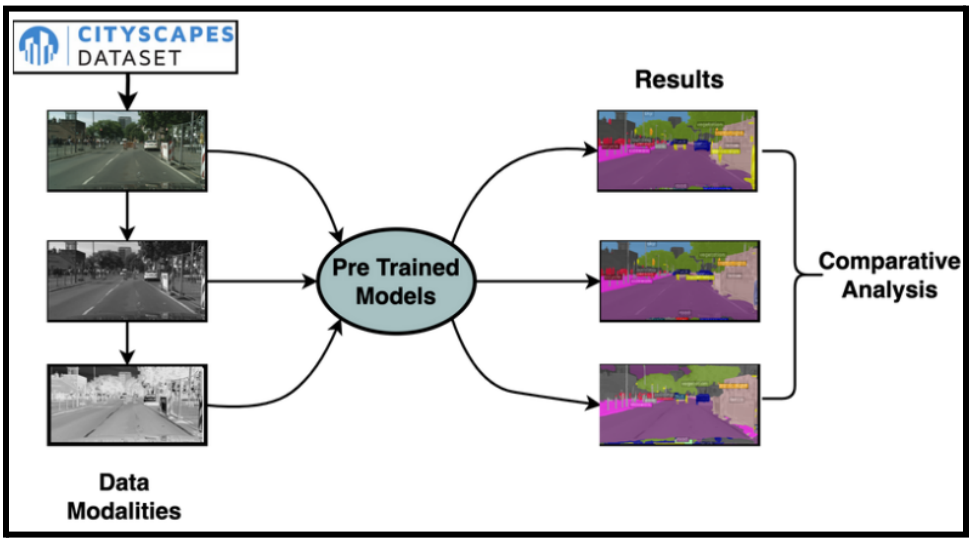
Assessing these models' capabilities involves a comparison across three distinct modalities: original RGB images, grayscale images, and inverted grayscale images. This evaluation strategy aims to elucidate how these models perform across different data modalities, representative of diverse scenarios. Specifically, these modalities serve as proxies for various real-world applications, including medical imaging such as X-rays (grayscale), inverted X-rays (grayscale inverted) and low-light conditions (grayscale).

The evaluation metric employed for comparing these models is the mean Intersection over Union (mIoU), a widely accepted measure in semantic segmentation tasks. mIoU is calculated by dividing the intersection of predicted and ground truth masks by the union of these masks across all classes, then averaging these values. By utilizing mIoU, we can objectively quantify and compare the models' performance across the diverse data modalities, providing a standardized evaluation criterion. This structured methodology enables a systematic and comprehensive evaluation, shedding light on the models' strengths and weaknesses, facilitating an informed analysis of biases and performance variations across diverse data modalities.

Experimentation

Dataset

Our exploration centered on the Cityscape dataset[15], a robust benchmark featuring high-resolution images of urban scenes. Cityscapes presents a diverse collection of urban street scenes captured across various



cities, meticulously annotated with pixel-level semantic labels for a wide range of classes. This extensive labeling provides a valuable resource for evaluating algorithms and models in the context of urban scene understanding. Of particular note, the pre-trained models under scrutiny were originally trained on Cityscapes dataset's RGB images. This initial training serves as a baseline performance measure, allowing for a comprehensive assessment of model capabilities on this specific data distribution. The dataset's richness in urban context and semantic annotations makes it a pivotal resource for research and development in computer vision applications related to urban environments. We utilized a subset of the dataset comprising a total of 5000 images with fine annotations. Specifically, 2975 images were allocated for training, 500 images for validation, and 1525 images for testing. For evaluation, we exclusively employed the 500 validation images, as we leveraged pre-trained models. The subset dataset encompasses annotations for 19 classes.

Utilizing the validation set from Cityscapes, we engaged the original RGB images as the primary source for evaluation. In addition to these original RGB images, we transformed the data into grayscale representations and further inverted these grayscale images. This conversion process facilitated the creation of three distinct data modalities for assessment. Each pre-trained model was tested on these diverse data distributions from the Cityscapes validation set, allowing us to comprehensively analyze their performance.

We initiated our experiments with DeepLabV3+ model which incorporates an encoder-decoder structure. The encoder employs dilated convolutions at multiple scales to process multiscale contextual information, while the decoder refines segmentation outcomes along object boundaries. In the process, encoder features undergo a 4x bilinear upsampling and are then concatenated with corresponding low-level features from a ResNet101 backbone, specifically extracted from the conv4_block6_2_relu block, optimizing the model's segmentation performance.

For SETR, we employed ViT-Large as the encoding backbone, and for decoding, PUP was chosen based on superior performance observed on RGB images compared to other decoding architectures. Additionally, MLA was used as a decoder for experiments conducted on inverted grayscale images. Instead of directly upscaling the output of the vision transformer to match the input image, PUP (Progressive UPscaling) carries out this process in multiple steps. MLA (Multi-Level feature Aggregation), akin to PVT, aggregates the transformer output at multiple levels, with the dimensions of each output being identical. The aggregated features are subsequently upscaled to match the original image size before being passed through the classifier.

The third model on which we conducted experiments was Segformer. In this case, we employed MIT-B5 as the encoding backbone. For decoding, Segformer aggregates features from various layers of PVT. Following the rescaling of each output, a straightforward MLP layer is utilized as a classifier.

Results & Analysis

In this section we will compare the three models for the three modalities. We will discuss our findings and explain their reasoning to the best of our knowledge, with supporting experiments if available.

Method	**Encoder**	**mIoU**	
		ADE20K	**Cityscapes**
DeeplabV3+	ResNet-101	44.1	80.9
SETR(PUP)	ViT-Large	50.2	82.2
SegFormer	MiT-B5	51.8	84.0

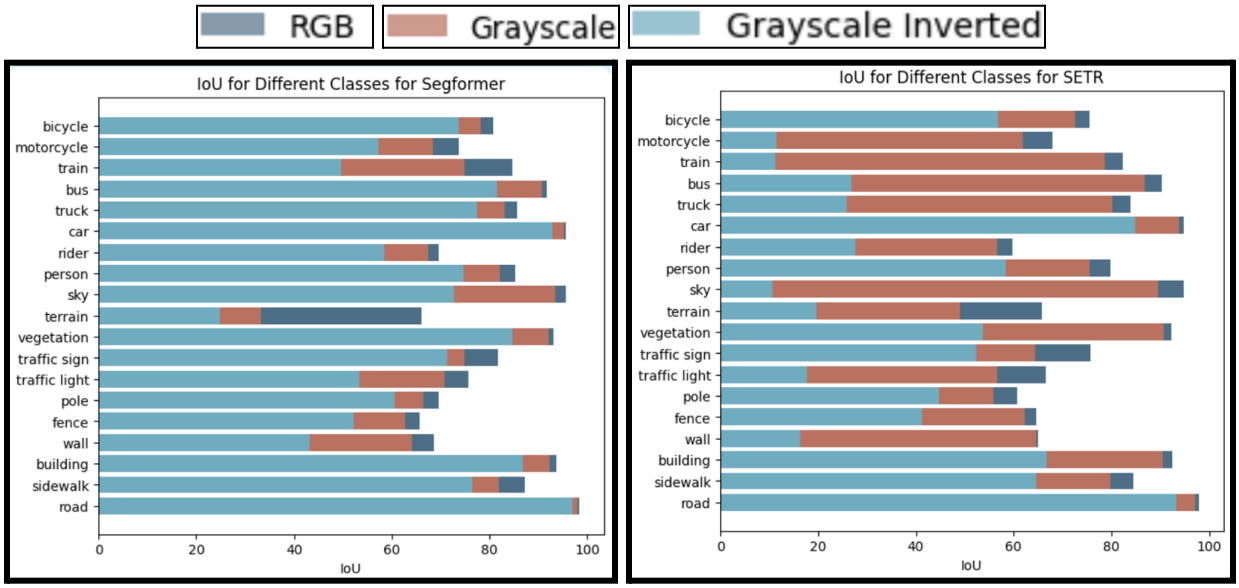
Table 1 compares the performance of DeeplabV3+, SETR and SegFormer on the datasets ADE20k and Cityscapes val set for the task of Image Segmentation. SegFormer performs best on both sets with achieving 84% mIoU on Cityscapes and 51.8% mIoU on ADE20k. SETR(PUP) (with Progressive UPsampling decoder) achieves mIoU of 82.2% and 50.2% on Cityscapes and ADE20k respectively. DeeplabV3+ with the backbone of Resnet-101 scores mIoU of 80.9 on Cityscapes and 44.1 on ADE20k datasets.

	DeeplabV3+	**SETR**	**SerFormer**
RGB	76.21	78.60	82.25
Grayscale	68.66	74.07	77.42
Grayscale Inverted	31.99	41.27	67.84

Table 2 compares all the models across all three data domains on Cityscapes validation dataset. When the RGB images are converted to Grayscale, all three models observe a slight decrease in performance. Majority of this performance can be attributed to a few images which contain a terrain class of labels. When

converted to grayscale, all three models start to miss-classify terrain as either vegetation or sidewalk. We hypothesize that this occurs due to the fact that the two major characteristics of the terrain class are its color, which it shares with vegetation, and its position on an image, which it shares with the sidewalk. Figure below reports mIoUs of individual classes for SETR and SegFormer. As we can absorb from Grayscale comparison, only the terrain class of the label observes major degradation.

An interesting observation can be made when Grayscale images are inverted to obtain inverted Grayscale images. When images are inverted, the magnitude of gradient remains the same, but the actual pixel intensities get mirrored about the intensity value 128. This process is detrimental for the task of classification due to the fact that the boundaries which differentiate different classes change drastically. Hence, we initially predicted that all three will see a sharp decrease in performance. To our surprise, this process didn't prove to affect SegFormer as we initially predicted. DeeplabV3+ saw a 58% decrease in performance when compared to RGB images. SETR performed a little better with only 47.5% relative performance degradation. We attribute this difference to global position embeddings used in the case of SETR. SegFormer only saw a performance decrease of 17.5%. When compared to other models, SegFormer



performed substantially better. After some investigation and ablation experiments on SETR (PUP vs MLA decoder), we found that SegFormer is robust due to it employing local as well as global attention. SegFormer only uses restricted attention heads, which calculates attention only in local areas. To calculate attention at different levels, SegFormer scales the images after certain transformer encoder blocks. SegFormer Decoder has information about high level coarse features as well as low level fine features. What this essentially means is that, while classifying a segment of image, DeeplabV3+ and SETR uses global information and does one-vs-all classification, SEgFormer uses local features to do one-vs-few classification. Since, the SegFormer also has position information, it has information about the local demographic between different classes, which it can use at the decoder level to reduce the classification problem from one-vs-all to one-vs-few.

Conclusion

In summary, SegFormer performs the best among DeeplabV3+, SETR, and itself across different data modalities. When comparing the models on different types of images (RGB, Grayscale, and Inverted Grayscale) in the Cityscapes dataset, all models show a slight decrease in performance with Grayscale images. However, SegFormer stands out by maintaining better performance when the Grayscale images are inverted. This resilience is attributed to SegFormer's unique use of local and global attention mechanisms, making it more effective in handling diverse image characteristics. Overall, SegFormer appears as a robust and effective model for image segmentation tasks across various scenarios. Additionally, it is crucial to exercise caution when fine-tuning pre-trained models on datasets with different distributions than those on which they were originally pre-trained, as these models inherently carry biases from their pre-training.

References

[1] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, *34*, pp.12077-12090.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[3] Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P. and Shao, L., 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 568-578).

[4] Islam, M.A., Jia, S. and Bruce, N.D., 2020. How much position information do convolutional neural networks encode?. *arXiv preprint arXiv:2001.08248*.

[5] Ronneberger, O., Fischer, P. and Brox, T., 2015. 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 (pp. 234-241). Springer International Publishing.

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

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

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

[9]Tsang,Sik-Ho,Medium, Review: DeepLabv3+ — Atrous Separable Convolution (Semantic Segmentation) | by Sik-Ho Tsang , Accessed 19th October, 2023

[10] Chen, T., Frankle, J., Chang, S., Liu, S., Zhang, Y., Carbin, M. and Wang, Z., 2021. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16306-16316)

[11] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248–255).

[12] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H. and Zhang, L., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with

transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6881-6890).
[13] Xie, Y. and Richmond, D., 2018. Pre-training on grayscale imagenet improves medical image classification. In *Proceedings of the European conference on computer vision (ECCV) workshops* (pp. 0-0).
[14] De, K. and Pedersen, M., 2021. Impact of colour on robustness of deep neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 21-30).
[15] Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., 2015, June. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision* (Vol. 2). Sn.
[16] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and 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), pp.834-848.
[17] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L., 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.
[18] Shyni, H.M. and Chitra, E., 2022. A comparative study of X-ray and CT images in COVID-19 detection using image processing and deep learning techniques. Computer Methods and Programs in Biomedicine Update, 2, p.100054.

Teamwork

The collaborative effort involved our team meeting at a shared location, where we collectively divided responsibilities. Three teammates took charge of working on one model each—Amogh Agashe worked on the SETR model, Rishabh Patil on Segformer, and Ameya Zope on DeepLabV3+. Preet Amin focused on dataset cleaning and processing, transforming data from RGB to grayscale and inverted grayscale. Additionally, Preet provided assistance to teammates encountering errors while working on the models. Following the model performance evaluations on different modalities, we collaborated on analysis and visualization. Formulation of the hypothesis was a joint effort, involving extensive discussions and research. Alongside creating the poster, all four team members actively participated in the analysis and hypothesis discussions, making it a collective endeavor without specific delineation of individual contributions.