JigsawGAN: A Novel Approach to Self-Supervised Image Segmentation

Raquel Peña Alarcón
Student Number - 23083963

University College London, London, UK
raquel.alarcon.23@ucl.ac.uk

1 Introduction

1.1 Background and Literature Review

The advent of supervised learning has significantly enhanced machine learning applications across various domains. However, it frequently faces a significant challenge: the requirement for extensive labeled datasets. This issue is particularly acute in fields like medical imaging, where obtaining labeled data can be prohibitively expensive. As a robust alternative, self-supervised learning (SSL) has emerged. This approach leverages unlabeled data to pre-train models, which are later fine-tuned using smaller labeled datasets. Combining the broad capabilities of unsupervised learning with the precision of supervised learning, SSL potentially reduces the dependency on large amounts of labeled data. Recent advancements have illustrated that models can effectively learn to recognize objects by analyzing spatial relationships within images. This is achieved by extracting random pairs of patches from each image and training a convolutional neural network (CNN) to predict the position of one patch relative to the other [4].

Generative Adversarial Networks (GANs) have significantly contributed to SSL's development, especially using tasks like puzzle-solving to improve learning depth and feature understanding without extensive labeled data [5, 2]. Studies such as those by Jin and Tian highlight the effectiveness of SSL tasks like jigsaws in improving segmentation and classification in GAN architectures [6]. This integration promises greater data efficiency and enhanced performance from limited datasets.

1.2 Motivation

Inspired by the pioneering work of Chen [3] which leverages the synergy of adversarial GAN learning and self-supervision for image synthesis without the need for labeled data, we are motivated to use this method. Additionally, the findings of Jing and Tian, advocating for the effectiveness of generative self-supervised techniques in learning image features, serving as pretext tasks for subsequent segmentation, all without relying on annotated labels, encouraged us to explore context-based methods such as Jigsaw puzzles. Finally, this same study led us to consider the integration of CNNs for fine-tuning, harnessing their sophisticated architectures designed for computer vision tasks [6] .

1.3 Research Questions

In this study, we aim to enhance segmentation tasks by integrating generation and context learning through self-supervised training of a GAN with a jigsaw puzzle-solving task. Our investigation focuses on comparing the image segmentation performance for cats and dogs of fine-tuned models against baseline ones without pre-training. Open ending questions driving our research include assessing the impact of the jigsaw puzzle-solving task by comparing models trained with and without it, as well as exploring the importance of information learned from the self-supervised task by evaluating segmentation performance with frozen pre-trained weights. These inquiries guide our exploration of self-supervised approaches and their role in improving image segmentation outcomes.

2 Methodology

The method presented in Figure 1 outlines a novel approach to self-supervised image segmentation, employing a generative adversarial network enhanced by jigsaw shuffling, followed by an optimization phase using CNN.

Pre-training Generator G accepts noise vector z, sampled from a normal distribution to generate synthetic images X_{fake}, X_{real} represents the samples from the original Animal-90 data distribution. Both sets undergo a shuffling *Jigsaw* transformation to obtain S_{fake} and S_{real} respectively. The Jigsaw Shuffler divides images into either 4 or 9 pieces to form a pretext task, then permutes these pieces using a selected set of 30 permutations based on Hamming distance to enrich the model's ability to discern. $\text{HD}(p1, p2)=\sum_{i=1}^{n} 1(p1_i \neq p2_i)$ where $p1$ and $p2$ are permutation indices of arranged tiles and it computes the number of permutation different. Discriminator D classifies X_{real} and X_{fake} as real and fake. For the discriminator, the objective is to minimize the error between the true shuffling order and the prediction for shuffling order of S_{real}, the reason for updating D is to learn inclusively and solely the features for real data; for the generator, the objective is to minimize the error between the true shuffling order and the prediction for S_{fake}. The loss *function*:

$$L_D = p \log(D(x_r)) + q \log(1 - D(x_f))$$

$$L_G = p \log(D(x_f)) + q \log(1 - D(x_r))$$

These equations reflect the binary cross entropy loss to measure the performance of the Discriminator and Generator; and also, another loss is calculated by the Jigsaw method. Weights are iteratively updated based on the gradient of the combined losses $w \leftarrow w - \eta \cdot \nabla L$ [3].

Fine-Tuning with Convolutional Neural Networks. Building on the pre-training described earlier, the fine-tuning process utilizes the already established generative architecture to specifically tailor it for image segmentation tasks. During this phase, the pre-trained weights of the discriminator D^5 are employed to initiate detailed feature extraction layers crucial for segmentation,

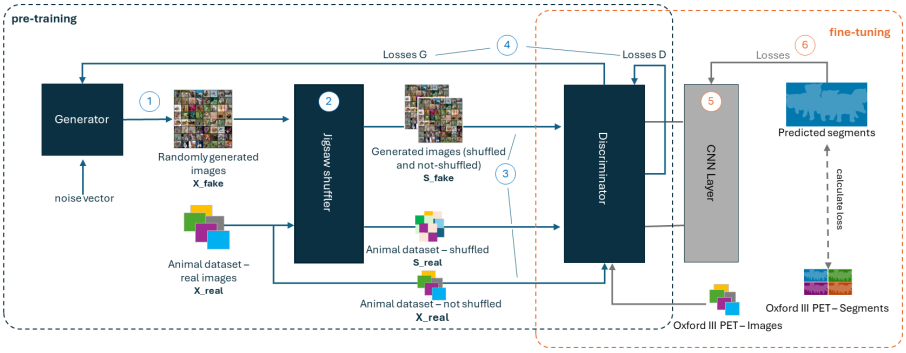

Fig. 1: Self-Supervised Jigsaw GAN Architecture. Dark blue lines indicate the pre-training process, incorporating shuffling with respective backpropagation paths. Gray lines delineate the fine-tuning phase with the CNN layer.

Fig. 2: Example shuffled images with 9 tiles

while transposed convolutional layers are used to progressively increase the resolution from the smallest dimensions to the desired final image size. Additionally, the loss[6] functions are dynamically updated during this phase to optimize both the discrimination accuracy and the segmentation precision; this process can be conducted with or without freezing the discriminator, freezing the weights allows the newly added layers to adapt freely without altering the previously learned representations. The final layers utilize an activation function to classify each pixel into relevant categories such as pets, alongside the refinement from the Jigsaw task for enhanced contextualization. This fine-tuning phase employs the Oxford-IIIT Pet dataset [7] to effectively train the model for high-precision segmentation tasks.

Metrics. The metrics used to evaluate the experiments include Multiclass Pixel Accuracy, which assesses pixel-wise classification accuracy across image channels; Intersection over Union (IoU) to gauge the accuracy of boundary delineation of the segmented objects; and CrossEntropyLoss, which measures the discrepancy between the predicted probabilities and the actual binary labels.

3 Experiments and Results

This section details the experimental setup used to evaluate the performance of our models:

1. Downstream comparison of the self-supervised GAN and JigsawGAN with a fully supervised baseline CNN for image segmentation on the Oxford-IIIT Pet Dataset.
2. Analysis of the effectiveness of integrating Jigsaw puzzles within the GAN architecture compared to using a standard GAN setup.
3. Ablation study to assess the impact of reducing the fine-tuning data size on model performance for the three models: JigsawGAN, GAN and CNN.

3.1 Experimental SetUp

Dataset. For the pre-training stage, we utilized the Animal Image Dataset from Kaggle [1], which consists of 5,400 unlabeled images representing 90 different animal species. This dataset was selected for its diversity, allowing for potential model scalability to recognize a wide range of pets, such as lizards and horses, and to facilitate differentiation among various animal species. For fine-tuning and testing, we employed the Oxford-IIIT Pet Dataset [7], which includes 37 pet breed categories totalling 7,349 images (2,371 of cats and 4,978 of dogs). This dataset is structured into three categories: object, edge, and background.

Baseline CNN. For the fully supervised model, we employ a CNN that integrates standard `Conv2d` layers for feature extraction and `ConvTranspose2d` layers for spatial upsampling, effectively localizing and segmenting within the Oxford-IIIT Pet dataset [7]. Training parameters include a cross-entropy loss function optimized via the Adam algorithm, with a learning rate of 0.0004 and a batch size of 16 across 20 epochs. The inclusion of `BatchNorm` and `ReLU` activations within the model enhances its capability to handle non-linearities effectively. The model employs `CrossEntropyLoss` for training optimization and utilizes Pixel Accuracy and IoU metrics for performance evaluation, with the Oxford-IIIT Pet dataset [7] split into 80% for training and 20% for validation.

Fine-Tunning. The JigsawGAN's Generator synthesizes images at a reduced resolution of 64×64 pixels, employing 5 `ConvTranspose2d` layers followed by a `ReLu` activation for normalization across 200 epochs. In contrast, the Discriminator uses 5 `Conv2d` layers to compress the spatial dimensions of input images followed by a `LeakyReLu` activation. The fine-tunning removes the this last Conv2d layer and attached the CNN with the `ConvTranspose2d` layers and 20 epochs. The Animals-90 dataset serves for pre-training, wherein images are randomly permuted in either 4 or 9 segments as an **additional experiment**; this is followed by fine-tuning and validation with the Oxford-IIIT Pet dataset [7]. The Adam optimizer, starting with a learning rate of 0.0002 for pre-training, is adaptable during fine-tuning. Binary Cross-Entropy Loss gauges image authenticity for the Discriminator, while Cross-Entropy Loss is employed for Jigsaw puzzle classification tasks. Pixel Accuracy and IoU metrics evaluate the model's performance during fine-tuning.

Ablation Study. This study was designed to assess the impact of reducing the training data size on the performance of both the JigsawGAN and the complete Baseline model. Fine-tuning was conducted using a range of data subsets, starting with 100% of the training set and decreasing by 10% increments down

to 10%. Cross-Entropy Loss and metrics like Pixel Accuracy and IoU are used to evaluate performance across each data size increment.

3.2 Summary Results

Model	Accuracy	IoU
Baseline	80.79%	59.08%
Jigsaw GAN (9 pieces)	81.99%	61.30%
Jigsaw GAN (4 pieces)	**82.19%**	**62.33%**
Non-jigsaw GAN	81.85%	60.18%
Frozen Weights		
Jigsaw GAN (9 pieces)	79.67%	56.15%
Jigsaw GAN (4 pieces)	80.10%	57.64%
Non-jigsaw GAN	79.76%	56.63%

Table 1: Segmentation performance, obtained from fine-tuning on Oxford-IIIT Pet dataset

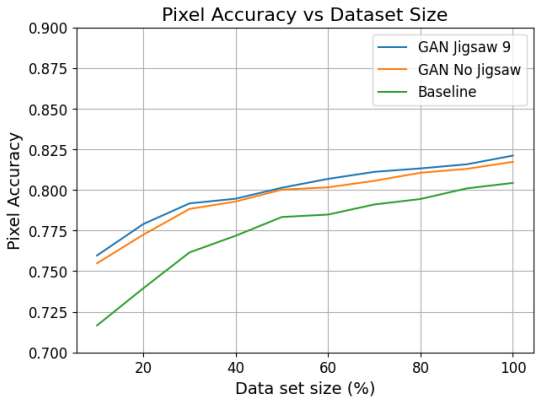

Fig. 3: Comparison of pixel accuracy across different dataset sizes

Downstream Performance. The 4-piece configuration of JigsawGAN demonstrates a marginal improvement in test performance, achieving 82.19% accuracy and 62.33% in IoU, thus surpassing the baseline CNN model and the GAN method. This comparison is conducted with unfrozen weights.

Efficacy of Jigsaw Integration. When contrasting JigsawGAN 9 pieces against a standard GAN setup without jigsaw, it displayed a slight increase in accuracy by 0.14% and a more pronounced improvement in IoU by 1.13%. This uptick suggests that Jigsaw may instill a more robust discriminatory capacity in the model, especially regarding edges, as evidenced by IoU. According to Fig. 5, it is observed that the loss on the training and validation set for JigsawGAN shows an initial divergence which stabilizes over time, demonstrating that it is learning effectively and not falling into overfitting.

Ablation Study on Data Size. Beyond 80% of the dataset size, no significant improvements in pixel accuracy are observed, indicating a stabilization in additional learning. The number of segments used for pre-training does not substantially affect outcomes on the validation set. Moreover, JigsawGAN consistently outperforms the Baseline model, emphasizing the advantage of self-supervisory tasks.

4 Discussion and Future Work

The findings from our study contribute valuable insights and demonstrate that JigsawGAN surpasses a fully supervised baseline model, signaling the potential

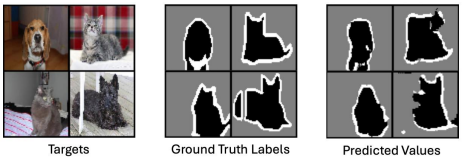

Fig. 4: Mask prediction for JigsawGAN with 9 pieces.

The Fig.4 present a qualitative assessment. The predicted labels approximate the ground truth with notable fidelity, particularly capturing the core outlines of the pet subjects. Minor deviations can be observed around complex edge details, suggesting areas for further refinement in the model's ability for border delineations.

of self-supervised learning to capture semantic information. Notably, the model appears to reach an optimal balance by providing enough complexity for learning without the overwhelming fragmentation seen in the 9-piece configuration. Furthermore, the improvement of GAN models over the baseline reaffirms the benefits of pre-training in generating and learning from augmented data. Contrary to expectations, freezing the weights during fine-tuning proved to be detrimental, as evidenced by reduced pixel accuracy and IoU across all GAN variations. This underscores the importance of dynamic weight adjustments during fine-tuning to allow the model to properly adapt to the nuances of image segmentation.

Further analysis of dataset complexity is warranted. The diversity of the Animal 90-Pet dataset facilitated preliminary feature learning; focusing solely on specific species such as cats and dogs might refine our understanding of GAN performance [1, 2]. Additionally, incorporating self-supervised tasks like rotation and translation could further enhance model robustness. Exploring the role of transposed layers and optimizing image resolution in tasks such as JigsawGAN might lead to more efficient feature extraction and improved segmentation outcomes

5 Conclusion

This work acknowledges prior research that has utilized the combination of GANs and Jigsaw tasks for shuffling and unshuffling images as a method of self-supervised training. This approach enhances the discriminator's capabilities and the overall quality of image generation without the need for class labels. However, our research introduces a significant innovation by incorporating fine-tuning with CNNs, which represents a notable advancement in the field of image segmentation.

We observed that simpler 4-piece jigsaw configurations outperform more complex setups in learning and generalization, evident when compared to baseline CNN models and traditional GANs without Jigsaw tasks. Our results also highlight that freezing weights during fine-tuning reduces performance, emphasizing the need for dynamic weight adjustments. By combining jigsaw-solving with adversarial training, we improve segmentation performance and model robustness significantly across various sample sizes compared to baselines.

References

[1] Sourav Banerjee. *Animal Image Dataset (90 Different Animals)*. `https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals`. 2023.

[2] Gulcin Baykal and Gozde Unal. *DeshuffleGAN: A Self-Supervised GAN to Improve Structure Learning*. 2020. arXiv: `2006.08694`.

[3] Ting Chen et al. "Self-Supervised GANs via Auxiliary Rotation Loss". In: (2019). arXiv: `1811.11212v2`.

[4] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. "Unsupervised Visual Representation Learning by Context Prediction". In: (2016). arXiv: `1505.05192v3`.

[5] Ian J. Goodfellow et al. "Generative Adversarial Networks". In: (2014). arXiv: `1406.2661`.

[6] Longlong Jing and Yingli Tian. "Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey". In: (2019). arXiv: `1902.06162`.

[7] O. M. Parkhi et al. "Cats and Dogs". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2012.

6 Appendix

6.1 Visualizations

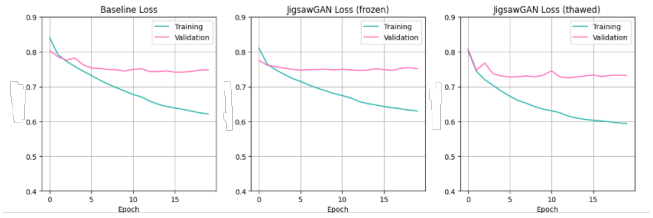

Fig. 5: Cross-Entropy Loss across epochs