DUAL-SPACE AUGMENTED INTRINSIC-LORA FOR WIND TURBINE SEGMENTATION

Shubh Singhal[1,2,*], *Raül Pérez-Gonzalo*[1,3,*], *Andreas Espersen*[3] *and Antonio Agudo*[1]

[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2]Université de Toulon, La Garde, France
[3]Wind Power LAB, Copenhagen, Denmark

ABSTRACT

Accurate segmentation of wind turbine blade (WTB) images is critical for effective assessments, as it directly influences the performance of automated damage detection systems. Despite advancements in large universal vision models, these models often underperform in domain-specific tasks like WTB segmentation. To address this, we extend Intrinsic LoRA for image segmentation, and propose a novel dual-space augmentation strategy that integrates both image-level and latent-space augmentations. The image-space augmentation is achieved through linear interpolation between image pairs, while the latent-space augmentation is accomplished by introducing a noise-based latent probabilistic model. Our approach significantly boosts segmentation accuracy, surpassing current state-of-the-art methods in WTB image segmentation.

Index Terms— Latent-space Augmentation, Diffusion Models, LoRA, Image Segmentation, Wind Turbine Blade.

1. INTRODUCTION

Operational damages to wind turbine blades (WTBs) can greatly impact their efficiency [1] and may even lead to complete failure [2]. Regular visual inspections and preventive maintenance are crucial to ensure timely repairs. These inspections are typically conducted using drones that capture high-resolution images, allowing for detailed analysis to guide maintenance decisions [3]. As the wind energy sector rapidly expands, the need for automated WTB assessment solutions is increasing, with image segmentation emerging as a key image processing task in this process [4].

Deep learning methods, particularly convolutional neural networks (CNNs), have driven significant advances in image segmentation research. Encoder-decoder architectures [5, 6] have become foundational frameworks by effectively capturing and reconstructing spatial relationships. Some notable models like DeepLabv3+ [7] and ResNeSt [8] achieve remarkable success by utilizing atrous convolutions and multiscale feature extraction techniques. The integration of attention mechanisms with CNNs, such as U-NetFormer [9], have further enhanced segmentation capabilities by improving global context understanding [10, 11]. In the realm of WTB segmentation, these advancements have inspired tailored models like BU-Net [4], which incorporates a post-processing hole-filling algorithm to refine segmentation results.

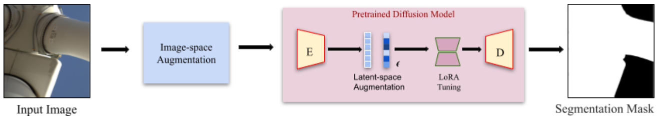

Fig. 1. **General schema of Segmentation-based Intrinsic LoRA (SI-LoRA) with dual-space augmentation.**

There has been growing interest in universal image segmentation models trained in a zero-shot manner using vast amounts of data, particularly large vision models like SAM [12] and DINO [13]. These models employ self-supervised vision transformers that learn meaningful representations from data through self-attention mechanisms. However, in practical applications, these universal segmentation methods often underperform compared to state-of-the-art models and cannot be trained directly due to the lack of comprehensive datasets. Intrinsic LoRA [14] presents a promising approach by fine-tuning generative models trained on large datasets, enabling supervised learning with minimal labeled data.

In this work, we extend Intrinsic LoRA for image segmentation and demonstrate its effectiveness in a real-world application. Specifically, we adapt pretrained Stable Diffusion models [15] for WTB segmentation by modifying the segmentation masks to meet the dimensional requirements of the pretrained model. These initial results, however, produce suboptimal performance. To address this, we explore several augmentation techniques. Initially, we apply traditional data augmentation methods to the input images [16, 17, 18], which prove particularly effective. Then, inspired by prior research that stabilizes the training of generative models in the latent space [19, 20, 21], we introduce a Bayesian adaptation of Intrinsic LoRA for image segmentation, modeling the latent vectors in a probabilistic augmented framework. By integrating both image-level and latent-space augmentations (see Fig. 1), our dual-space augmentation approach substantially improves segmentation performance, surpassing state-of-the-art methods in WTB segmentation by a large margin.

This work has been supported by the Innovation Fund Denmark under 2021 ID1044-0044A, by the project GRAVATAR PID2023-151184OB-I00 funded by MCIU/AEI/10.13039/501100011033 and by ERDF, UE; and by the project GreenVAR of the Fundación Ramón Areces. *Equal contribution.

2. METHODOLOGY

This section outlines our adaptation of the Intrinsic LoRA [14] method for image segmentation. We begin by reviewing the core principles of Intrinsic LoRA and its application in extracting image intrinsics. Next, we proceed by illustrating how we tailored this approach to create segmentation maps. Finally, we propose a novel dual-space augmentation method that operates in both image and latent spaces.

2.1. Intrinsic LoRA

Intrinsic-LoRA [14] harnesses the implicit understanding of image intrinsics within generative models to produce high-quality supervised outputs. By introducing learnable LoRA [22] adaptors $\boldsymbol{\theta}$, an image-to-image generative diffusion model can be fine-tuned with minimal labeled samples to generate the desired outputs $\mathbf{y} \in \mathbb{R}^{H \times W \times 3}$.

Given a pretrained Stable Diffusion model [15], the input image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ is encoded by the encoder E to a lower-dimensional latent space $\mathbf{z}_{\mathbf{x}}^{(E)} = E(\mathbf{x})$. The obtained latent vector $\mathbf{z}_{\mathbf{x}}^{(E)}$ is fed to the denoising U-Net [23] model $U_{\boldsymbol{\theta}}$ along with a text prompt t. This prompt is the image intrinsic to be extracted like ”depth”, ”normal” and so forth, and is encoded by a pretrained CLIP [24] tokenizer T, obtaining the transformed output latent vector $\mathbf{z}_{\mathbf{x}}^{(U)} = U_{\boldsymbol{\theta}}(\mathbf{z}_{\mathbf{x}}^{(E)}, T(t))$.

Intrinsic-LoRA adapts the diffusion model to a supervised task by optimizing the LoRA adaptors on top of the self- and cross-attention layers [25] of a single step dense predictor U-Net model. The adaptors $\boldsymbol{\theta}$ are optimized to minimize the differences between the transformed latent vector $\mathbf{z}_{\mathbf{x}}^{(U)}$ and the encoded ground-truth $\mathbf{z}_{\mathbf{y}}^{(E)} = E(\mathbf{y})$:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}}[d(\mathbf{z}_{\mathbf{x}}^{(U)}, \mathbf{z}_{\mathbf{y}}^{(E)})] \,, \tag{1}$$

where d is a specific-task dissimilarity metric. Finally, the decoder D transforms back $\mathbf{z}_{\mathbf{x}}^{(U)}$ to the image space, obtaining the predicted intrinsic map $\hat{\mathbf{y}} = D(\mathbf{z}_{\mathbf{x}}^{(U)})$. Both the encoder E and decoder D are frozen during training.

2.2. Segmentation-based Intrinsic LoRA (SI-LoRA)

Intrinsic LoRA is built upon image-to-image generative models, thus, it handles 3-channel inputs and outputs. However, in our image segmentation problem, we need to distinguish between background and foreground, requiring a single-channel output. Hence, we define $\mathbf{y}$ as the concatenation along the third dimension of the ground-truth segmentation mask $\mathbf{m} \in \mathbb{R}^{H \times W}$. Similarly, the model's decoded output $\hat{\mathbf{y}}$ remains a 3-channel image, which we convert back into a single-channel mask $\hat{\mathbf{m}}$ by averaging across the three channels $\hat{\mathbf{y}}_c$, obtaining

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}}[\mathrm{MSE}(\mathbf{z}_{\mathbf{x}}^{(U)}, \mathbf{z}_{\mathbf{y}}^{(E)})], \quad \mathbf{y} = \mathbf{m} \otimes \mathbb{1}_3, \quad \hat{\mathbf{m}} = \frac{1}{3}\sum_{c=1}^{3} \hat{\mathbf{y}}_c \,, \tag{2}$$

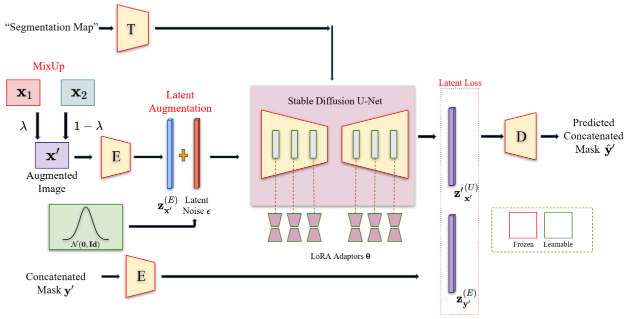
"Segmentation Map"
T
MixUp
$\mathbf{x}_1$
$\mathbf{x}_2$
λ
$1-\lambda$
$\mathbf{x}'$
Augmented Image
E
Latent Augmentation
$\mathbf{z}_{\mathbf{x}'}^{(E)}$
Latent Noise ϵ
$\mathcal{N}(\mathbf{0}, \mathbf{Id})$
Stable Diffusion U-Net
LoRA Adaptors $\boldsymbol{\theta}$
Latent Loss
$\mathbf{z}'^{(U)}_{\mathbf{x}'}$
$\mathbf{z}_{\mathbf{y}'}^{(E)}$
D
Predicted Concatenated Mask $\hat{\mathbf{y}}'$
Frozen
Learnable
Concatenated Mask $\mathbf{y}'$
E

Fig. 2. **Segmentation-based Intrinsic LoRA (SI-LoRA) architecture with dual-space augmentation (DSA).**

where $\mathbb{1}_3$ is a 3-dimensional vector of ones, and $\otimes$ denotes the outer product. Additionally, two adjustments were made to ensure effective segmentation: the dissimilarity metric d between latent vectors is measured using Mean Squared Error (MSE), and the text prompt t is set to ”segmentation map.”

2.3. Dual-space Augmentation (DSA SI-LoRA)

After successfully adapting the Intrinsic-LoRA method for our segmentation task, we shift our focus to enhance its performance. Data augmentation techniques in the image space have been widely explored and implemented to improve learning-based models [17, 18]. One particularly notable method is MixUp [16], which generates synthetic images through the linear interpolation of multiple samples. Specifically, given two images and their corresponding labels from the training set, $(\mathbf{x}_1, \mathbf{m}_1)$ and $(\mathbf{x}_2, \mathbf{m}_2)$, MixUp produces a new augmented sample $(\mathbf{x}', \mathbf{m}')$ as follows:

$$\mathbf{x}' = \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \,, \quad \mathbf{m}' = \lambda \mathbf{m}_1 + (1-\lambda)\mathbf{m}_2 \,, \tag{3}$$

where $\lambda \in [0, 1]$ is a mixing coefficient sampled from a Beta distribution with parameters $\alpha, \beta = 0.4$, which governs the interpolation between the two images and their labels.

While these techniques have proven effective, our contribution lies in extending augmentation to the latent space. Traditional diffusion models operated in the image space, however, Stable Diffusion demonstrated the effectiveness of shifting the diffusion process to the latent space [15]. Drawing inspiration from this and VAEs [26], we augment the training by parametrizing the latent vector $\mathbf{z}_{\mathbf{x}'}^{(E)}$ as a Bayesian input for the U-Net $U_{\boldsymbol{\theta}}$. In particular, the augmented $\mathbf{z}'^{(E)}_{\mathbf{x}'}$ is modeled as an isotropic Gaussian with an identity covariance matrix:

$$\mathbf{z}'^{(E)}_{\mathbf{x}'} \sim \mathcal{N}\big(E(\mathbf{x}'), \mathbf{Id}\big) \,. \tag{4}$$

This probabilistic approach enhances the U-Net's robustness by accommodating a wide range of latent inputs, instead of relying solely on deterministic encodings. To manage the stochastic nature of the sampling during backpropagation, we reparameterize the sampling process to a fixed base distribution. Consequently, the augmented latent input $\mathbf{z}'^{(E)}_{\mathbf{x}'}$ is computed by introducing a noise variable $\boldsymbol{\epsilon}$, drawn from a standard multivariate Gaussian distribution:

$$\mathbf{z}'^{(E)}_{\mathbf{x}'} = E(\mathbf{x}') + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{Id}) \,. \tag{5}$$

An overview of Segmentation-based Intrinsic LoRA (SI-LoRA) with dual-space augmentation is illustrated in Fig. 2.

3. EXPERIMENTAL RESULTS

In the following section, we first present the implementation details used to successfully train our model, along with the dataset employed. Next, we provide an in-depth evaluation of the performance of SI-LoRA, including each data augmentation strategy. This is followed by qualitative assessments that highlight the effectiveness of our dual-space augmented SI-LoRA. We then compare our model with various state-of-the-art segmentation algorithms. Finally, we demonstrate its robustness by comparing its performance across different windfarms in the test set, showcasing exceptional results across diverse environments.

3.1. Dataset and Implementation Details

The dataset utilized to train (1712 images) and evaluate (320 images) the proposed method is taken from [4]. The input images and ground-truth segmentation masks are resized to 512×512. Decoupled regularization [27] with a weight decay of 10^{-2}, an initial learning rate of 10^{-4} and a batch size of 2 is employed. The training is stopped after 30 epochs. For LoRA adaptors, we choose the rank 8, consistent with the original study [14]. For generating binary masks, we use Otsu's method to threshold the model predictions. The experiments were performed on an NVIDIA GeForce RTX 3090.

3.2. Ablation Study

Ablation studies were conducted to better understand the individual contribution of different augmentation techniques in the image and latent space. We compare four different model configurations: (1) SI-LoRA (Sec. 2.2) without data augmentation, (2) SI-LoRA with image-space augmentation implemented in terms of MixUp [16], (3) SI-LoRA with latent-space augmentation implemented in terms of noise-based probabilistic model (Sec. 2.3), and (4) SI-LoRA with both image- and latent-space augmentation. Distinct metrics are evaluated to highlight the contributions of each data augmentation strategy, including the overall performance metrics of accuracy, recall, F1-score, and mean IoU (mIoU).

Tab. 1 showcases that applying no augmentation techniques (row 1) results in the lowest performance across all metrics, serving as a baseline for comparison. Introducing latent-space augmentation alone (row 2) shows a significant improvement in all metrics, particularly a 22.03% increase in F1-score and a 20.48% improvement in mIoU. This suggests that augmenting the latent space helps the model generalize better by simulating diverse, realistic variations in feature representations. When only image-space augmentation is applied (row 3), the model further boosts performance, effectively enriching the training data with more variability. Finally, combining both augmentation strategies (row 4) yields the highest performance across all metrics, with an accuracy of 99.15%, F1-score of 98.84%, and an mIoU of 97.69%. These results demonstrates the synergistic benefits of applying both image- and latent-space augmentation techniques in the SI-LoRA framework. While each augmentation method contributes independently to model performance, their combined effect leads to the most significant improvements across all evaluated metrics.

Table 1. Dual-space augmentation ablation study.

MixUp [16]	Latent Noise	Accuracy (%)	Recall (%)	F1 (%)	mIoU (%)	Relative F1 (%)	Relative mIoU (%)
No	No	82.30	73.86	76.48	76.15	100.00	100.00
No	Yes	95.22	91.93	93.33	91.76	122.03	120.48
Yes	No	97.59	96.75	97.23	95.24	127.13	125.05
Yes	Yes	**99.15**	**98.60**	**98.84**	**97.69**	**129.24**	**128.28**

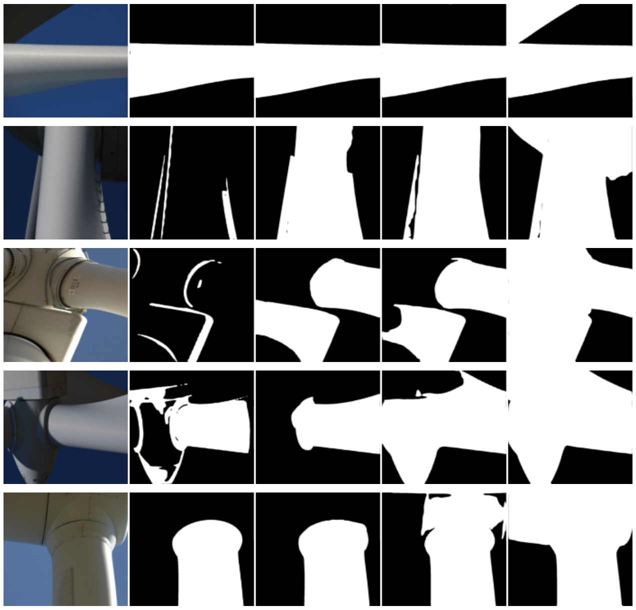

Fig. 3. Qualitative comparison of distinct data augmentation strategies. From left to right: Input image, SI-LoRA (Sec. 2.2), SI-LoRA with image-space augmentation (MixUp [16]), SI-LoRA with latent-space augmentation (Sec. 2.3), and SI-LoRA using both augmentations.

3.3. Qualitative Evaluation

To qualitatively illustrate the segmentation masks produced by various augmentation strategies, Fig. 3 presents five examples where dual-space segmentation is essential for achieving high-quality results. As discussed in Sec. 3.2, SI-LoRA struggles to generate smooth maps, leading to poor segmentation in many instances. While SI-LoRA can detect some edges of the wind turbine structure (see the second and third examples in Fig. 3), it fails to identify all the turbine components, focusing only on a single section.

When applying image- or latent-space augmentation, we observe that SI-LoRA produces smoother maps, and the generated masks include the inner regions of the wind turbine blade (WTB), rather than just outlining the structure's edges. However, even with this improvement, SI-LoRA still falls short of fully capturing all parts of the wind turbine, as seen in all five instances. By combining both augmentation strategies, the model effectively resolves these challenging cases,

Table 2. Quantitative comparison with competing models.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	mIoU (%)	IoU_{bckg} (%)	$\text{IoU}_{\text{blade}}$ (%)
SW [6]	93.48	93.57	91.71	91.37	87.44	88.64	86.23
DeepLabv3+ [7]	94.14	96.36	87.38	89.03	87.47	90.31	84.62
ResNeSt [8]	94.23	96.84	91.47	92.77	89.63	90.40	88.86
SAM [12]	94.36	97.29	91.22	92.60	91.66	92.31	91.01
DiffSeg [28]	96.37	83.20	89.74	85.73	86.40	91.67	81.13
U-NetFormer [9]	96.20	97.31	93.51	94.42	91.75	92.53	90.96
BU-Net [4]	97.39	**99.42**	93.35	95.73	93.80	94.70	92.90
SI-LoRA (Sec. 2.2)	82.30	97.01	73.86	76.48	76.15	79.56	72.74
DSA SI-LoRA (Sec. 2.3)	**99.15**	99.20	**98.60**	**98.84**	**97.69**	**97.56**	**97.83**

generating highly accurate segmentation masks. In particular, it successfully captures all parts of the wind turbine, including darker regions on a secondary plane, as evident in the first and second examples. These results demonstrate the effectiveness of our approach in handling complex segmentation scenarios.

3.4. Quantitative Evaluation

To evaluate our proposed method, we conducted a comparative analysis as shown in Tab. 2, benchmarking SI-LoRA against popular segmentation models. The evaluation was performed on a test set of 200 turbine images from various windfarms [4]. In its initial form, SI-LoRA underperformed across all metrics, lagging behind all competing models. This was primarily due to overfitting on the training masks, which significantly hindered its ability to generalize to newly acquired, unseen test images. This limitation prompted us to explore augmentation techniques aimed at enhancing its performance and generalization capabilities.

By introducing dual-space augmentation (DSA), combining MixUp [16] for image-space variability with noise-based probabilistic models for latent-space diversification, we developed DSA SI-LoRA. As shown in the table, DSA SI-LoRA dramatically outperforms the original SI-LoRA, surpassing state-of-the-art models by a large margin across all major metrics, except precision. These results highlight the effectiveness of our augmentation strategies in enhancing generalization and overall performance, overcoming overfitting and improving segmentation models' robustness in real-world applications. This demonstrates that pretrained generative models can be efficiently fine-tuned with limited data to perform real-world supervised tasks, such as WTB segmentation.

3.5. Windfarm Dissimilarity

The test dataset used in our study comprises 20 images from various windfarms [4], captured using different drone configurations and locations. To evaluate the robustness of dual-space augmented SI-LoRA (DSA SI-LoRA), Fig. 4 presents a boxplot illustrating the performance across 10 distinct windfarms. This figure offers insights into the robustness of DSA SI-LoRA in WTB image segmentation.

The boxplot reveals consistently high average performance metrics, including accuracy, F1-score, and mIoU, with minimal variability in performance distribution. These results indicate that DSA SI-LoRA effectively generates accurate masks across a range of input environments.

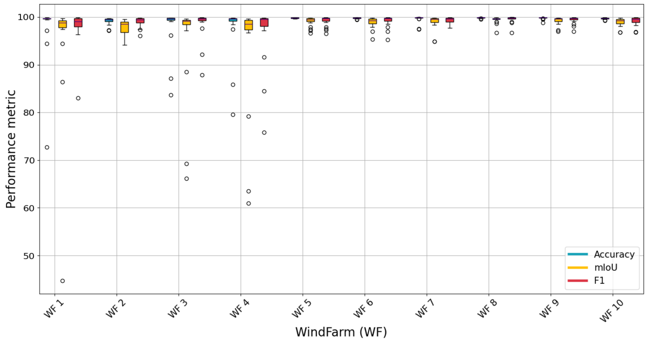

Fig. 4. Boxplot test results across different windfarms.

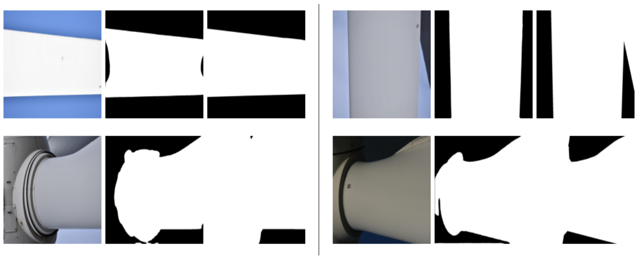

Fig. 5. Failure Cases. From left to right: Input Image, SI-LoRA using both augmentations (Sec. 2.3), and ground-truth segmentation masks. On both sides of the figure, the same information is displayed.

Nevertheless, the boxplot also highlights a few outliers, with lower performance observed for windfarms 1, 3, and 4. Fig. 5 displays four representative examples where the method encounters difficulties. In these instances, high contrast between the WTB and the background, or within the WTB region itself, can lead to parts of the WTB being misclassified as background. Despite these isolated cases, they do not detract significantly from the overall robustness demonstrated by the method.

4. CONCLUSION

In conclusion, this paper presents a significant advancement in wind turbine blade (WTB) image segmentation through the development of the dual-space augmented Segmentation-based Intrinsic LoRA (SI-LoRA). By extending the capabilities of Intrinsic LoRA to image segmentation and employing an innovative dual-space augmentation strategy, our method fine-tunes generative pretrained models using minimal data, addressing the limitations of large vision universal models in specialized domains. In particular, the dual-space strategy integrates linear interpolation in the image space and probabilistic augmentation in the latent space, leading to substantial improvements in segmentation accuracy. Our experiments demonstrate that dual-space augmented SI-LoRA consistently outperforms existing state-of-the-art models in WTB segmentation, delivering robust performance across windfarms. These results highlight the potential of SI-LoRA as a powerful tool for improving the automation and reliability of wind turbine maintenance, ultimately contributing to the sustainability and efficiency of wind energy operations.

5. REFERENCES

[1] P. Haselbach, R. Bitsche, and K. Branner, “The effect of delaminations on local buckling in wind turbine blades,” *Renewable Energy*, vol. 85, pp. 295–305, 2016.

[2] Y. Lin, L. Tu, H. Liu, and W. Li, “Fault analysis of wind turbines in china,” *RSER*, vol. 55, pp. 482–490, 2016.

[3] R. Pérez-Gonzalo, A. Espersen, and A. Agudo, “Generalized nested latent variable models for lossy coding applied to wind turbine scenarios,” in *ICIP*, 2024.

[4] R. Pérez-Gonzalo, A. Espersen, and A. Agudo, “Robust wind turbine blade segmentation from rgb images in the wild,” in *ICIP*, 2023, pp. 1025–1029.

[5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE TPAMI*, vol. 40, no. 4, pp. 834–848, 2017.

[6] X. Pan, X. Zhan, J. Shi, X. Tang, and P. Luo, “Switchable whitening for deep representation learning,” in *ICCV*, 2019, pp. 1863–1871.

[7] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *ECCV*, 2018, pp. 801–818.

[8] H. Zhang et al., “Resnest: Split-attention networks,” in *CVPRW*, 2022, pp. 2736–2746.

[9] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, “Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery,” *JPRS*, vol. 190, pp. 196–214, 2022.

[10] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, and H. Hu, “Disentangled non-local neural networks,” in *ECCV*, 2020, pp. 191–207.

[11] Y. Meng, H. Zhang, D. Gao, Y. Zhao, X. Yang, X. Qian, X. Huang, and Y. Zheng, “BI-GConv: Boundary-aware input-dependent graph convolution for biomedical image segmentation,” in *BMVC*, 2021.

[12] A. Kirillov et al., “Segment anything,” in *ICCV*, 2023, pp. 4015–4026.

[13] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *CVPR*, 2021, pp. 9650–9660.

[14] X. Du, N. Kolkin, G. Shakhnarovich, and A. Bhattad, “Intrinsic lora: A generalist approach for discovering knowledge in generative models,” in *CVPRW*, 2024.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022, pp. 10684–10695.

[16] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” in *ICLR*, 2018.

[17] D. Walawalkar, Z. Shen, Z. Liu, and M. Savvides, “Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification,” in *ICASSP*, 2020, pp. 3642–3646.

[18] J. Zhang, Y. Zhang, and X. Xu, “Objectaug: object-level data augmentation for semantic image segmentation,” in *IJCNN*, 2021, pp. 1–8.

[19] C. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, “Amortised map inference for image super-resolution,” in *ICLR*, 2017.

[20] S. Jenni and P. Favaro, “On stabilizing generative adversarial training with noise,” in *CVPR*, 2019, pp. 12145–12153.

[21] F. Zhu, Z. Cheng, X.-y. Zhang, and C.-l. Liu, “Class-incremental learning via dual augmentation,” *NeurIPS*, vol. 34, pp. 14306–14318, 2021.

[22] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” in *ICLR*, 2022.

[23] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *MICCAI*, 2015, pp. 234–241.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., “Learning transferable visual models from natural language supervision,” in *ICLR*, 2021, pp. 8748–8763.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017.

[26] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *ICLR*, 2014.

[27] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *ICLR*, 2019.

[28] J. Tian, L. Aggarwal, A. Colaco, Z. Kira, and M. Gonzalez-Franco, “Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion,” in *CVPR*, 2024, pp. 3554–3563.