

DAKD: Data Augmentation and Knowledge Distillation using Diffusion Models for SAR Oil Spill Segmentation

Jaeho Moon^{1*}Jeonghwan Yun^{1*}Jaehyun Kim^{1*}Jaehyup Lee^{2†}Munchurl Kim^{1†}¹Korea Advanced Institute of Science and Technology²Kyungpook National University<https://kaist-viclab.github.io/DAKD-site>

Abstract

Oil spills in the ocean pose severe environmental risks, making early detection essential. Synthetic aperture radar (SAR) based oil spill segmentation offers robust monitoring under various conditions but faces challenges due to the limited labeled data and inherent speckle noise in SAR imagery. To address these issues, we propose (i) a diffusion-based Data Augmentation and Knowledge Distillation (DAKD) pipeline and (ii) a novel SAR oil spill segmentation network, called SAROSS-Net. In our DAKD pipeline, we present a diffusion-based SAR-JointNet that learns to generate realistic SAR images and their labels for segmentation, by effectively modeling joint distribution with balancing two modalities. The DAKD pipeline augments the training dataset and distills knowledge from SAR-JointNet by utilizing generated soft labels (pixel-wise probability maps) to supervise our SAROSS-Net. The SAROSS-Net is designed to selectively transfer high-frequency features from noisy SAR images, by employing novel Context-Aware Feature Transfer blocks along skip connections. We demonstrate our SAR-JointNet can generate realistic SAR images and well-aligned segmentation labels, providing the augmented data to train SAROSS-Net with enhanced generalizability. Our SAROSS-Net trained with the DAKD pipeline significantly outperforms existing SAR oil spill segmentation methods with large margins.

1. Introduction

Oil spills, caused by offshore drilling, shipping accidents, and natural seepage, lead to the release of vast amounts of oil into oceans, posing severe, often irreversible threats to marine ecosystems and coastal economies. This requires early detection and accurate damage assessment. Synthetic Aperture Radar (SAR) stands out among remote sensing modalities, as it acquires data in all weather and light-

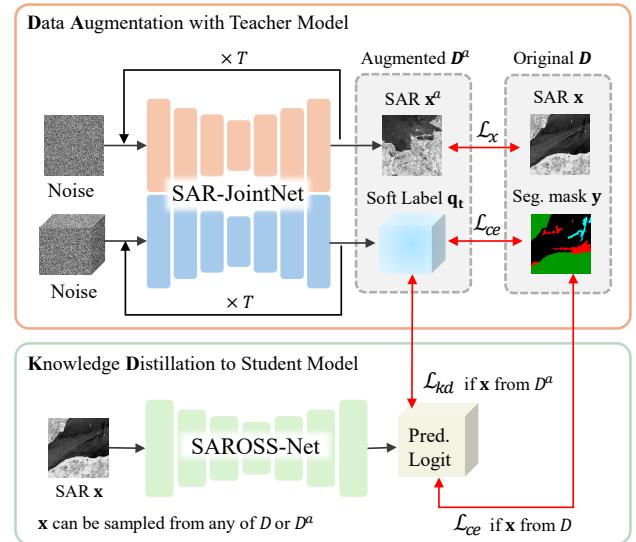


Figure 1. Overview of our Data Augmentation and Knowledge Distillation (DAKD) pipeline. The diffusion-based SAR-JointNet learns to generate SAR images and their soft labels for data augmentation and knowledge distillation to our student SAROSS-Net.

ing conditions by emitting microwave signals and measuring the reflected backscatter from the Earth's surface [16, 21, 35, 38]. So, the analysis of SAR imagery is essential for the detection and estimation of oil spill areas in the ocean. Additionally, oil spills result in reduced surface roughness, causing the affected areas to appear as dark spots in SAR images [54, 63]. By detecting these dark regions, SAR-based methods can accurately identify and delineate the extent of an oil spill [26, 48, 56].

However, a major challenge in SAR-based oil spill segmentation is in two folds: (i) The lack of oil spill data and their precise labels - SAR oil spill data is limited due to the rarity of oil spill events and the complexity of label annotation; (ii) The inherent speckle noise in SAR images - this is caused from random interference among elementary reflectors, which obscures essential features and complicates the differentiation between targets and background [43, 57, 64].

*Co-first authors, equal contribution;

†Co-corresponding authors.

These limitations underscore the difficulties of applying traditional segmentation models to SAR imagery, especially for oil spill detection.

In this paper, to tackle the scarcity of labeled SAR oil spill data, we *firstly* propose a diffusion-based Data Augmentation and Knowledge Distillation (DAKD) technique for oil spill segmentation. The success of generative models, particularly diffusion models [31], has led to their increasing use in data augmentation [1, 23, 36, 40, 68, 69]. Here, we further extend the capability of diffusion models to knowledge distillation by generating SAR images and logit values for the corresponding labels for knowledge distillation. To achieve this, we propose a SAR-JointNet which jointly models the distribution of SAR images and segmentation maps by accounting for different information retention in the diffusion process across modalities. Also, to faithfully produce pixel-wise class probabilities (soft label) with rich information from logits, the label generation is optimized using a cross-entropy loss. By leveraging the augmented data with soft labels from the SAR-JointNet, our teacher model, we can induce robust training for the student network, called SAR Oil Spill Segmentation Network (SAROSS-Net). Compared to one-hot encoded hard labels, these soft labels, containing richer probability information for each class, allow SAROSS-Net to learn more effectively and improve segmentation performance.

Unlike natural images, SAR imagery contains excessively abundant high-frequency noise, making selective feature extraction among noisy features critical for precise segmentation. To address this, we introduce Context-Aware Feature Transfer (CAFT) blocks in the skip connections to selectively transfer the encoded features from the MambaVision encoder [28] to the decoder for our SAROSS-Net. Our CAFT block facilitates effective referencing of semantic features to prioritize essential high-frequency features, enhancing the segmentation accuracy under noisy conditions.

Fig. 1 illustrates our DAKD pipeline. Our SAR-JointNet is trained to generate SAR images and their corresponding logits for segmentation using the original training dataset \mathcal{D} . After training, our SAR-JointNet synthesizes an augmented dataset \mathcal{D}^a including the SAR images and their logits. Leveraging both \mathcal{D} and \mathcal{D}^a , we train our proposed SAROSS-Net with knowledge distillation from SAR-JointNet. Our SAROSS-Net trained with the DAKD pipeline demonstrates superior performance on the SAR oil spill datasets [37, 81], achieving notable improvements in accuracy and robustness compared to previous approaches. Our contributions are summarized as follows:

- To the best of our knowledge, we *firstly* propose a Data Augmentation and Knowledge Distillation (DAKD) pipeline, which leverages a diffusion-based teacher model trained for data generation to augment data (DA) and enables knowledge distillation to a student model (KD).

- By leveraging SAR-JointNet that can effectively learn the joint distribution of the SAR images and segmentation mask for data augmentation, we handle the data scarcity in SAR ‘oil-spill’ segmentation.
- We introduce a SAR oil-spill segmentation network, SAROSS-Net, with Context-Aware Feature Transfer (CAFT) blocks in skip connections, to selectively transfer high-frequency features from noisy SAR images to the decoder blocks for better segmentation accuracy.
- Our pipeline significantly outperforms existing state-of-the-art methods on the SAR oil spill detection dataset [37] and the SOS dataset [81].

2. Related Work

2.1. Semantic Segmentation on SAR images

With the rise of deep learning, semantic segmentation in natural images has advanced based on convolutional neural network (CNN) architectures [2, 9, 41, 45, 75] and attention mechanisms [19, 20, 24, 51, 62]. Recently, foundation models like SAM [33, 34] have demonstrated innovative zero-shot semantic segmentation capabilities by utilizing prompts such as points, bounding boxes, and text. In addition, Mamba-based architectures [28, 77] introduce a new encoder that combines CNNs for robust local feature extraction with Structured State Space Models [27] and self-attention [62] layers to capture both short- and long-range dependencies. However, these models still struggle to handle remote sensing modalities like SAR imagery [64].

In remote sensing, deep learning-based semantic segmentation models have also been studied for satellite Electro-Optical and SAR images [78–80]. Specifically, for SAR oil spill segmentation, various models such as U-Net, LinkNet, and Deeplab are widely used [3, 15, 37, 81]. Krestenitis *et al.* [37] utilized Deeplabv3+ [11] model on the oil spill detection dataset consisting of five classes such as ‘land’, ‘sea-surface’, ‘ship’, ‘look-alike’ and ‘oil spill’. In addition, Zhu *et al.* [81] proposed a SAR oil spill (SOS) dataset and a new architecture, CBD-Net. More recently, SAM-OIL [66] proposed a two-stage approach: first detecting oil spills with a YOLOv8 [18] network, then applying large-scale SAM [34] inference based on bounding boxes predicted from the first stage. However, the scarcity of publicly available oil spill data continues to hinder the development and generalization of these methods, limiting their robustness across different real-world scenarios.

2.2. Data Augmentation with Diffusion Models

Data augmentation increases the diversity and amount of data in situations with limited data to improve the performance of several tasks. Recent studies utilize the power of diffusion models to generate diverse images and corresponding label pairs for data augmentation. For image clas-

sification, [1, 25, 40, 61] used text as both the class label and the condition, generating image-class label pairs and improved performance. In image captioning, [17] addressed limited artistic data by generating image-caption pairs. Object detection benefits from models like [7, 23], which generate images based on bounding boxes, while [73] creates image-bounding box pairs jointly from diffusion features. For semantic segmentation, [55] generates image-scribble pairs to enhance scribble-supervised segmentation, while [69] generates medical image-segmentation mask pairs using masks as conditions. For remote sensing images, Sat-Synth [60] jointly synthesizes Electro-optical images and corresponding segmentation masks by simply using concatenated images and labels as inputs of a diffusion model.

Our diffusion-based SAR-JointNet synthesizes SAR images and segmentation masks to address the scarcity of labeled SAR oil spill data. In the training of the diffusion-based joint generation of images and labels, balancing the differences between modalities is critical for downstream tasks such as semantic segmentation. Our SAR-JointNet introduces a balancing factor to effectively model the joint distribution of SAR images and segmentation masks.

2.3. Knowledge Distillation

Knowledge distillation (KD) has emerged as a prominent technique for transferring information from a larger or more complex model (teacher model) to a smaller and simpler model (student model). Initially introduced by Hinton *et al.* [30], KD allows the student model to learn from the softened output predictions of the teacher, capturing both the correct class and the relative probabilities of other classes, which are often useful for learning generalizable features. KD has been actively studied [49, 74] and useful across a variety of computer vision tasks [39, 44, 65, 71].

KD has also been studied to address data dependency during the training of neural networks. The data-free knowledge distillation approaches [6, 22, 46] enable knowledge transfer from teachers to student models by generating synthetic data that mimic the training distribution, even when the original dataset is unavailable due to privacy concerns. Recently, diffusion-based distillation techniques leverage pretrained large-scale 2D diffusion models to transfer knowledge to new tasks such as 3D novel view synthesis [42, 50]. In addition, DiffKD [32] is proposed to align denoised student features to teacher features, enhancing the performance of image classification and semantic segmentation. In our proposed Data Augmentation and Knowledge Distillation (DAKD) pipeline, our diffusion-based SAR-JointNet not only learns to generate realistic SAR images and their corresponding labels that mimic the training data distribution but also provides logit values containing the teacher’s knowledge. This allows the student model to effectively learn from the augmented data.

3. Method

3.1. DAKD Pipeline

Our Data Augmentation and Knowledge Distillation (DAKD) framework to effectively handle the scarcity of labeled data for SAR oil spill segmentation is depicted in Figure 1. Our DAKD pipeline consists of two stages.

In the first stage of the DAKD pipeline, our SAR-JointNet is trained to generate an augmented dataset, \mathcal{D}^a , including SAR images along with their corresponding logits. Unlike one-hot segmentation maps (hard labels), logits (soft labels) provide probabilistic information that contains class predictions of our SAR-JointNet. This richer supervision can offer stable learning for the student model by leveraging pixel-level certainty, which helps accurate segmentation and strengthens the model’s generalization during training. Our SAR-JointNet is trained to ensure that the generated SAR images and labels closely resemble the original data distribution, allowing them to effectively supplement the existing dataset.

In the second stage, our SAROSS-Net, a SAR oil spill segmentation network, is trained using the combination of the original training dataset (\mathcal{D}) and the augmented dataset (\mathcal{D}^a) synthesized by our SAR-JointNet. From a knowledge distillation perspective, the synthesized soft labels in \mathcal{D}^a offer improved training stability and guide robust segmentation learning, compared to hard labels.

3.2. SAR-JointNet

Our SAR-JointNet takes advantage of the JointNet [72] that diffuses two modalities into their respective networks: (i) The usage of separate networks for two modalities (SAR images and segmentation masks in our case) enables the utilization of one single modality (SAR images only) to improve its generation ability even when the labels (segmentation masks) are unavailable; (ii) By making the two networks focus on their own modalities, they can be specialized in learning the unique features of their own. Therefore, in our SAR-JointNet, we implemented two networks: SAR-Net for SAR image generation and Label-Net for the corresponding logit generation. SAR-Net and Label-Net interact with each other to exchange their essential features across different modalities, enabling cross-modality reinforcement and well-aligned SAR and logit pair generation. Fig. 2 shows the overall architecture of SAR-JointNet consisting of SAR-Net and Label-Net.

3.2.1 Balancing Diffusion between Two Modalities

Diffusion-based joint generation of the natural images and their pixel-wise labels has been explored [60, 72], but the different characteristics across modalities are overlooked. In generating noisy SAR images x and segmentation masks

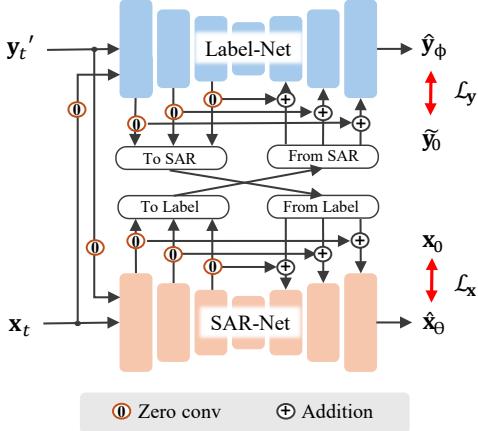


Figure 2. Overall architecture of our SAR-JointNet.

\mathbf{y} simultaneously, the information level in each modality can vary significantly. Therefore, balancing the retained information level between the two modalities is essential for stable training. To address this issue, Chen *et al.* [12, 14] applied a predefined scale factor to the images in the diffusion process to adjust the information level of noise corrupted images between different image resolutions or modalities.

We extend this approach by introducing a balancing factor b based on the signal-to-noise ratio (SNR). For input SAR $\mathbf{x}_0 \in \mathbb{R}^{H \times W \times 1}$ and the normalized segmentation masks $\mathbf{y}_0 \in \{-1, 1\}^{H \times W \times C}$ where C is the number of classes, b is multiplied to \mathbf{y}_0 , to control information levels between \mathbf{x}_t and \mathbf{y}_t , expressed as:

$$\mathbf{y}'_0 = b\mathbf{y}_0. \quad (1)$$

Using \mathbf{x}_0 and $\mathbf{y}'_0 \in \{-b, b\}^{H \times W \times C}$, forward diffusion process at timestep t is expressed as follows:

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon, \quad \mathbf{y}'_t = \alpha_t \mathbf{y}'_0 + \sigma_t \epsilon, \quad (2)$$

where α_t and σ_t are the degrees of signals and noise, respectively. The SNR of \mathbf{x}_t and \mathbf{y}'_t are formulated as follows:

$$\begin{aligned} \text{SNR}(\mathbf{x}_t) &= P(\alpha_t \mathbf{x}_0)/P(\sigma_t \epsilon) \\ &= \alpha_t^2 P(\mathbf{x}_0)/\sigma_t^2 P(\epsilon), \\ \text{SNR}(\mathbf{y}'_t) &= \alpha_t^2 b^2 P(\mathbf{y}_0)/\sigma_t^2 P(\epsilon) \\ &= b^2 \text{SNR}(\mathbf{y}_0), \end{aligned} \quad (3)$$

where $P(\cdot)$ denotes the mean power of frequency components, obtained from the 2D discrete FT. The balancing factor b that matches $\text{SNR}(\mathbf{x}_t)$ to $\text{SNR}(\mathbf{y}'_t)$ is defined as:

$$\begin{aligned} b &= \sqrt{\text{SNR}(\mathbf{x}_t)} / \sqrt{\text{SNR}(\mathbf{y}_t)} \\ &= \sqrt{P(\mathbf{x}_0)} / \sqrt{P(\mathbf{y}_0)}. \end{aligned} \quad (4)$$

In our experiments, we calculate b by averaging the mean power of the frequency component from the entire training dataset, which is shared across whole training samples.

3.2.2 Training Strategy of SAR-JointNet

Our SAR-JointNet leverages a three-stage training strategy, to progressively enhance the capability of SAR images and their labels (segmentation masks). In Stage 1 of training, SAR-Net θ is trained to denoise the noise-corrupted SAR image \mathbf{x}_t by minimizing \mathcal{L}_x between the original SAR image \mathbf{x}_0 and the denoised SAR image $\hat{\mathbf{x}}_\theta$, expressed as:

$$\mathcal{L}_x = \mathbb{E}_{\mathbf{x}_0} [\|\mathbf{x}_0 - \hat{\mathbf{x}}_\theta\|_2^2]. \quad (5)$$

In Stage 2 of training, Label-Net ϕ is initialized with the pretrained weights of SAR-Net from Stage 1, while the input and output layers are reinitialized to match the channel number of the labels. To enhance label generation capability, SAR-Net is frozen except for the layers interacting with Label-Net such as zero-convolution layers. To guide Label-Net training, we employ a cross-entropy (CE) loss that has been demonstrated to be effective for accurate label generation in prior work [13, 14]. Label-Net learns to generate logits (pixel-wise probability map) $\hat{\mathbf{y}}_\phi$, by minimizing \mathcal{L}_y , defined as:

$$\mathcal{L}_y = -\frac{1}{N} \sum_{j=1}^N \sum_{i=1}^C \tilde{\mathbf{y}}_0(i, j) \log(\hat{\mathbf{y}}_\phi(i, j)), \quad (6)$$

where $N = H \times W$ and $\tilde{\mathbf{y}}_0 \in \{0, 1\}^{H \times W \times C}$ denotes the one-hot encoded segmentation mask, while SAR-Net is trained by minimizing \mathcal{L}_x . The combined loss $\mathcal{L}_{\text{joint}}$ for SAR-JointNet training is formulated as:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_x + \mathcal{L}_y. \quad (7)$$

In Stage 3, all layers of our SAR-JointNet are finetuned with a reduced learning rate by minimizing $\mathcal{L}_{\text{joint}}$ in Eq. 7. After training, our SAR-JointNet jointly synthesizes realistic SAR images and their logit values for our DAKD pipeline, with the label inference process following the approach of Chen *et al.* [14].

3.3 SAR Oil Spill Segmentation Network

By leveraging the proposed DAKD pipeline, our SAR-JointNet transfers knowledge to our proposed SAR oil spill segmentation network (SAROSS-Net). Fig. 3 illustrates the architecture of our SAROSS-Net based on a MambaVision [28] encoder. A key to achieving accurate segmentation from noisy SAR oil spill images is to filter out noise features while retaining the features for segmentation boundaries. We introduce Context-Aware Feature Transfer (CAFT) blocks to selectively transfer high-frequency features from encoded features to decoder blocks. Along UNet++ [76] architecture-based skip-connection, each CAFT block transfers high-resolution (HR) features to the next CAFT block or the convolutional decoder block, by referring to the lower-resolution (LR) features.

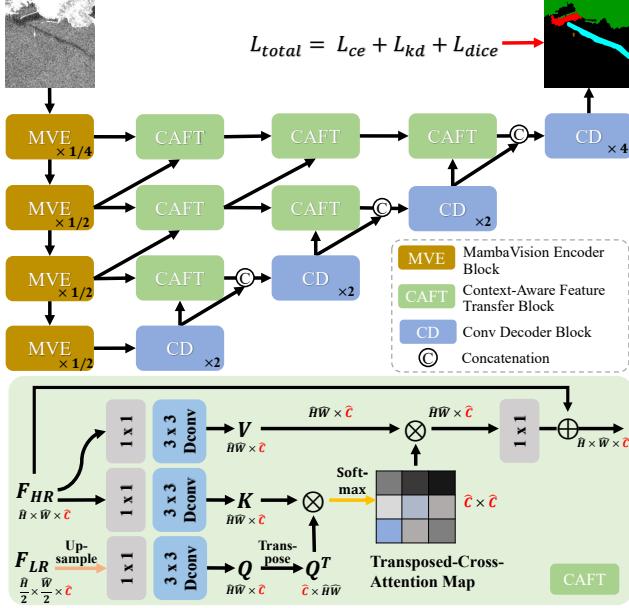


Figure 3. Overall architecture of our SAR Oil Spill Segmentation Network (SAROSS-Net).

3.3.1 Context-Aware Feature Transfer Block

As illustrated in Fig. 3, our proposed Context-Aware Feature Transfer (CAFT) block leverages the core principles of cross-attention, where LR feature maps contain semantic information as the Query, while HR feature maps as the Key and Value. This allows the CAFT block to effectively transfer only essential high-frequency spatial details from noisy SAR images guided by the abstract semantics from LR features. For the CAFT block, we adopt transposed-attention [70] to reduce the computational and memory burdens of the conventional attention mechanism [62]. By constructing the attention map in the channel dimension, the transposed-attention lowers the complexity from $O(H^2W^2)$ to $O(C^2)$, where $H \times W$ denotes the spatial resolution and C is the number of channels in the feature map. In our CAFT block, we use (i) the upsampled low-resolution features F_{LR} as Query features containing global semantic information, and (ii) the HR features F_{HR} as both Key and Value, capturing finer spatial details essential for precise segmentation. This enables the CAFT block to effectively integrate high-level semantics with detailed spatial information, enhancing the overall performance of the segmentation task.

3.3.2 Loss Functions for Training SAROSS-Net

To leverage the augmented SAR images and corresponding logits (\hat{y}_ϕ) from SAR-JointNet, we employ a knowledge distillation loss [30] to transfer knowledge to our SAROSS-

Net. For class i , the soft label $q_t(i)$ is defined as:

$$q_t(i) = \frac{\exp(\hat{y}_\phi(i)/T)}{\sum_{j=1}^C \exp(\hat{y}_\phi(j)/T)}, \quad (8)$$

where T is a hyperparameter that softens class probability distribution and C is the number of classes. Then, the knowledge distillation loss \mathcal{L}_{kd} is defined as a scaled sum of cross-entropy losses between soft label q_t and predicted logit p_s from our SAROSS-Net, which is expressed as:

$$\mathcal{L}_{kd} = \frac{T^2}{N} \sum_{j=1}^N \mathcal{L}_{ce}(q_t, p_s). \quad (9)$$

This knowledge distillation process allows our SAROSS-Net to learn from the richer and softened outputs (pixel-wise probability maps for predicted segmentation masks) of SAR-JointNet, improving the model’s generalizability.

Our SAROSS-Net is trained with a composite loss function including a cross-entropy loss \mathcal{L}_{ce} , a knowledge distillation loss \mathcal{L}_{kd} , and a soft dice loss \mathcal{L}_{dice} [59] that mitigates the class imbalance between foreground (oil-spilled sea surface) and background (clean sea surface). The total loss function is expressed as:

$$\mathcal{L}_{total} = \lambda_{ce} \mathcal{L}_{ce} + \lambda_{dice} \mathcal{L}_{dice} + \lambda_{kd} \mathcal{L}_{kd}, \quad (10)$$

where λ_{ce} , λ_{dice} , and λ_{kd} are empirically set as 1, 0.5, and 0.1, respectively.

4. Experiments

4.1. Datasets

Oil Spill Detection (OSD) dataset [37] was constructed using SAR images from the Sentinel-1 mission, collected via the European Space Agency (ESA) database and the Copernicus Open Access Hub. It consists of 1,002 images for training and 110 images for testing. Each image contains annotated instances across five semantic classes: ‘oil-spill’, ‘look-alike’, ‘ship’, ‘land’, and ‘sea surface’. The images were preprocessed through radiometric calibration, speckle noise filtering, and conversion to real luminosity values. Each SAR image has a resolution of $650 \times 1,250$.

Deep-SAR Oil Spill (SOS) dataset [81] contains SAR images obtained from two areas: the Gulf of Mexico captured by PALSAR sensor on the ALOS satellite, and the Persian Gulf captured by Sentinel-1A satellite. The dataset contains 8,070 labeled SAR images of 256×256 resolution, generated from 21 raw SAR images through techniques such as cropping, rotating, and adding noise to enhance data diversity. For the ALOS dataset, 3,101 images are allocated for training, and 776 images are used for testing. Similarly, for the Sentinel-1A dataset, 3,354 images are used for training, and 839 images are used for testing.

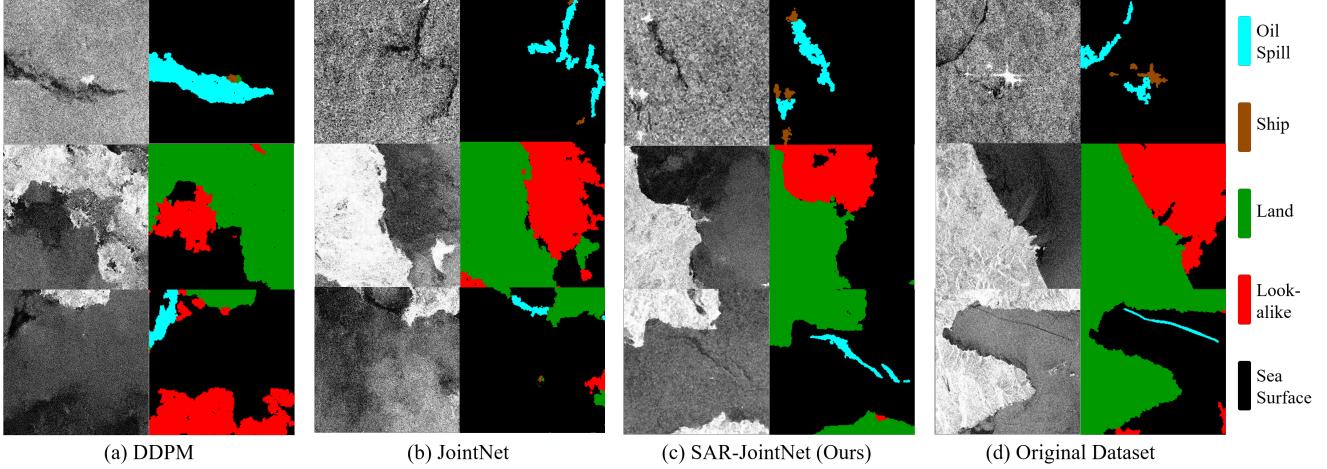


Figure 4. Joint generation results: Qualitative comparison of SAR images and corresponding segmentation masks generated by: (a) DDPM [31], (b) JointNet [72], and (c) SAR-JointNet (ours) to the samples from (d) the Original OSD dataset [37].

4.2. Implementation Details

Our SAR-JointNet is a pixel-domain diffusion model designed for realistic SAR image generation, whereas the original JointNet [72] is latent-based. We train SAR-JointNet in three stages. In Stage 1, we train SAR-Net for 50K iterations for both datasets [37, 81]. In Stage 2, SAR-JointNet is trained for 250K iterations on the OSD dataset [37] and 150K on the SOS dataset [81]. Lastly, in Stage 3, we finetune SAR-JointNet with 100K and 50K iterations on the OSD dataset [37] and the SOS dataset [81], respectively. We use AdamW optimizer [47] with a learning rate of $1e^{-4}$ for Stage 1 and 2, and $1e^{-5}$ for Stage 3. The training image and segmentation mask pairs are randomly cropped to a resolution of 256×256 on the OSD dataset [37]. We utilize a squared cosine DDPM beta scheduler with 1,000 steps for training, while we use the DDIM [58] scheduler with 200 steps for data generation. Training is conducted on two Nvidia A6000 GPUs with a batch size of 16 per GPU. The balancing factor b is set to 0.528 for OSD dataset [37], 0.447 for SOS-ALOS [81], and 0.466 for SOS-Sentinel [81]. For DAKD, we generate 4,480 and 6,400 pairs of image and labels, each in a resolution of 256×256 , for OSD [37], SOS dataset (ALOS & Sentinel) [81], respectively.

Our SAROSS-Net is based on the MambaVision-B encoder [28] and trained for 200 epochs with a batch size of 16, a learning rate of $5e^{-5}$, and the AdamW [47] optimizer. For the OSD dataset [37], input images are set to a resolution of 512×512 . From the original training dataset, SAR images are randomly cropped to 512×512 , while four randomly selected 256×256 synthetic SAR images are combined to form 512×512 , similar to the mosaic augmentation used in YOLOv4 [4]. For testing, the original resolution ($650 \times 1,250$) is used. For the SOS dataset [81], we train

	OSD - Sentinel [37]		SOS - Sentinel [81]		SOS - ALOS [81]	
	FID (↓)	IS (↑)	FID (↓)	IS (↑)	FID (↓)	IS (↑)
DDPM [31]	49.42	2.149	47.71	1.754	45.74	1.822
JointNet [72]	46.46	1.929	30.47	1.632	47.61	1.828
SAR-JointNet	30.64	2.186	19.88	1.809	21.23	1.948

Table 1. Quality comparison of generated SAR images between our SAR-JointNet and other diffusion-based generative models.

the SAROSS-Net with input images of 256×256 .

4.3. Data Generation Results from SAR-JointNet

We compare the generation qualities of SAR images and segmentation masks generated by our SAR-JointNet against other diffusion models, DDPM [31] and JointNet [72]. For training DDPM [31], we adopt the methodology from Sat-Synth [60], where SAR images and segmentation masks are simply concatenated as input. Our baseline model, JointNet [72], is trained without utilizing the balancing factor b and the cross entropy loss for label generation, making it a direct comparison point to highlight the improvements introduced by our approach. In Tab. 1, we compare the qualities of generated SAR images by measuring the Frechet inception distance (FID) and the inception score (IS) between SAR images from the original training datasets and augmented SAR images. For this analysis, we use 4,480 SAR images from the OSD dataset [37] and 3,200 from the SOS dataset [81]. Our SAR-JointNet achieves superior results in all metrics, especially with significantly large margins of 18.78 (38%) and 15.82 (32%) improvements in FID, compared to DDPM [31] and JointNet [72], respectively, for the OSD dataset [37].

Fig. 4 shows the generated SAR images and their corresponding segmentation masks (labels). In the first row, we

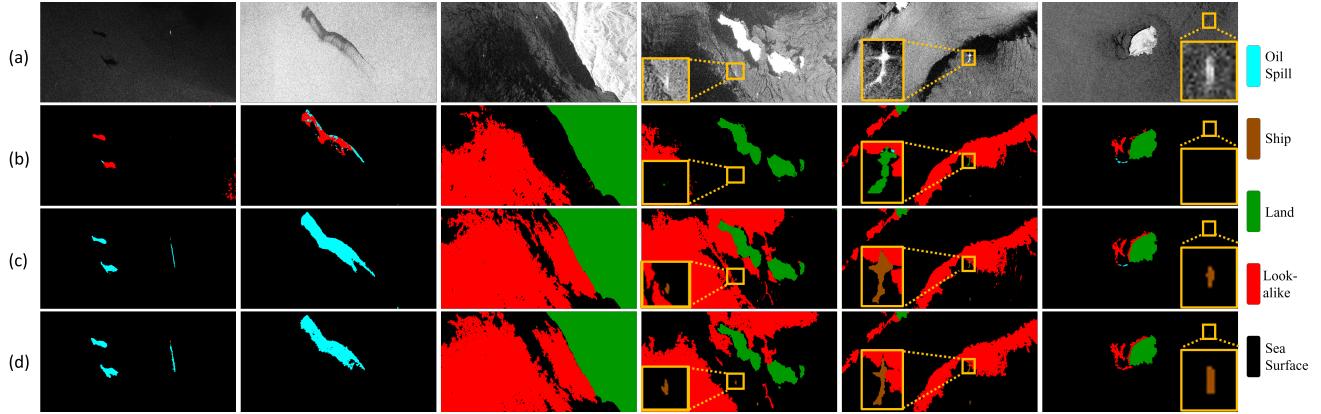


Figure 5. Oil spill segmentation results: Qualitative comparison of SAR oil spill segmentation on the OSD dataset [37]. (a) SAR Images, (b) CBD-Net [81], (c) SAROSS-Net (ours), and (d) Ground Truth.

compare generated SAR image and label pairs, containing ‘oil-spill’ and ‘ship’ class information. DDPM [31] misclassified the ship class with irrelevant land information, and JointNet [72] mismatched ship shapes. In contrast, our SAR-JointNet generates bright and star-shaped ships with precise segmentation mask alignment, closely matching the original training dataset. In the second row, DDPM [31] synthesizes unclear boundaries for the land, and JointNet [72] yields labels to include ‘look-alike’ (small-sized isolated red area) in the land region (green region). However, our SAR-JointNet accurately generates the SAR image with clear and accurate boundaries of the land and a precise segmentation mask. In the third row, DDPM [31] and JointNet [72] generate SAR and segmentation masks with poor correspondences, while SAR-JointNet successfully generates hazy dark areas representing ‘look-alike’, similar to the original datasets, and generates relatively distinct and elongated dark areas for ‘oil-spill’, which is not contained in the generated SAR images by the other two models.

4.4. SAR Oil Spill Segmentation Results

In Tab. 2, we compare the segmentation quality of our SAROSS-Net to other methods [8, 67, 81]. On the OSD dataset [37], our method consistently outperforms other approaches across all evaluation metrics, specifically achieving 11.66% mIoU improvement over SegFormer [67]. Also, for the SOS dataset, originally released by CBD-Net [81], our method achieves superior results for both ALOS and Sentinel-1A satellite images, with improvements of 2.99% and 3.61% in mIoU over CBD-Net [81]. This highlights the superiority of our method across different sensor modalities, further validating its robustness and effectiveness in various remote sensing scenarios.

Tab. 3 further provides the IoU scores for each class. Specifically, for the ‘sea surface’, ‘look-alike’, ‘oil-spill’, and ‘land’ classes, our method surpasses the previous SOTA

D	Methods	mIoU (%)	F1 (%)	A (%)	R (%)	P (%)
OSD	CBD-Net [81]	59.19	70.41	98.09	64.60	82.25
	DeepLabV3+ [11]	62.59	74.25	89.70	70.06	80.31
	SegFormer [67]	63.20	74.23	97.85	77.87	72.56
	Ours	74.86	84.27	99.03	83.24	85.50
SOS-ALOS	CBD-Net [81]	83.31	84.84	94.43	86.75	84.20
	DeepLabV3+ [11]	78.74	87.49	93.04	87.67	87.70
	SegFormer [67]	82.25	89.84	94.39	89.32	90.39
	Ours	86.30	92.23	95.65	92.28	92.18
SOS-Sent.	CBD-Net [81]	83.64	93.69	90.07	93.13	94.27
	DeepLabV3+ [11]	76.95	86.81	88.20	86.35	87.36
	SegFormer [67]	82.16	90.13	91.01	90.33	89.94
	Ours	87.25	93.05	93.68	93.24	92.88

Table 2. Oil spill segmentation performance comparison on OSD dataset [37] and SOS dataset [81] (ALOS & Sentinel-1A satellites). ‘D’ in the first column indicates datasets.

Methods	Sea	Oil	Look	Ship	Land	mIoU
UNet [53]	93.90	53.79	39.55	44.93	92.68	64.97
CBD-Net [81]	95.17	42.84	38.38	27.58	91.95	59.19
Segformer [67]	94.42	39.77	48.79	36.18	96.86	63.20
LinkNet [5]	94.99	51.53	43.24	40.23	93.97	64.79
PSPNet [75]	92.78	40.10	33.79	24.42	86.90	55.60
DeepLabv2 [10]	94.09	25.57	40.30	11.41	74.99	49.27
DeepLabv2 (msc) [10]	95.39	49.53	49.28	31.26	88.65	62.83
DeepLabV3+ [11]	88.43	44.80	56.44	31.06	92.21	62.59
SAM-OIL [66]	96.05	51.60	55.60	52.55	91.81	69.52
Ours	97.56	60.01	67.61	51.77	97.35	74.86

Table 3. Oil spill segmentation performance comparison on the OSD dataset [37], showing IoU (%) for 5 classes and mIoU (%).

method, SAM-OIL [66], with a significant improvement of 9.47% for the ‘oil-spill’ class. As shown in Fig. 5, CBD-Net does not effectively capture small objects such as ‘ship’, while our SAROSS-Net precisely identifies these objects in alignment with the ground truth. Additionally, our model accurately distinguishes ‘oil-spill’ regions, even those with shapes similar to ‘look-alike’ areas, demonstrating its ro-

CE loss	scale w/ b	FID (\downarrow)	IS (\uparrow)	mIoU (%)	F1 (%)
✓	✓	46.46	1.929	71.67	81.87
		44.81	2.158	74.08	83.76
	✓	32.28	2.206	72.65	82.77
✓	✓	30.64	2.186	74.86	84.27

Table 4. Ablation analysis on SAR-JointNet: FID and IS reflect SAR image generation quality on the OSD dataset [37], while mIoU and F1 scores denote segmentation performance when SAROSS-Net is trained on each augmented dataset.

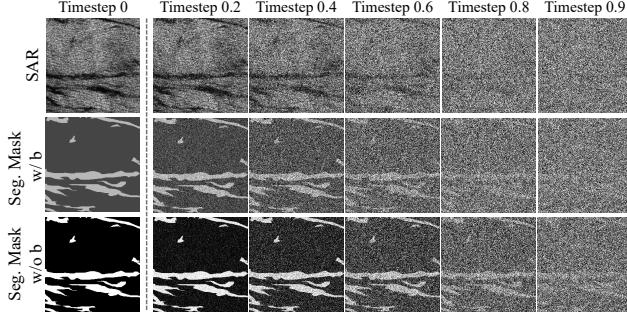


Figure 6. Effect of the balancing factor b (Eq. (4)). Noise-corrupted segmentation masks with applying the balancing factor contain a similar level of information to the noise-corrupted SAR images, compared to segmentation masks without applying it.

bustness in handling challenging classes.

5. Ablation Study

5.1. SAR Oil Spill Segmentation Network

Ablation studies were performed to show the effect of key components in our approach. First, we demonstrate the effect of employing the CE loss and the balancing factor b in Tab. 4 for training SAR-JointNet. With CE loss, SAR-JointNet provides soft label information that contains richer information than hard labels, yielding better segmentation performance in both mIoU and F1 scores, compared to the baseline model without CE loss. The balancing factor b effectively adjusts the information levels of masks with respect to their corresponding SAR images, leading to a significant improvement in both FID and IS for the generated SAR images, compared to the baseline. As shown in Fig. 6, the noise-corrupted segmentation masks scaled with the balancing factor maintain similar levels of information with the corresponding SAR image across all timesteps t . Especially at $t = 0.9$, the noise-corrupted segmentation mask without applying the balancing factor still contains the overall location of the oil spill, which can predefine the shape of the SAR image, potentially limiting the learning of joint distribution. Our SAR-JointNet, with both components applied, generates realistic SAR images and segmentation masks with strong correspondences, making it suitable for train-

Encoder	CAFT	HL	SL	mIoU (%)	F1 (%)	R (%)	P (%)	A (%)
ResNet [29]	✓			70.55	80.94	79.10	83.26	98.61
	✓	✓		72.33	82.36	80.85	84.60	98.90
	✓		✓	73.49	83.38	82.41	84.72	98.81
Mamba Vision [28]				73.34	84.19	83.08	84.66	98.91
				71.99	82.12	81.06	83.48	98.60
				73.38	83.25	81.93	84.70	98.82
	✓	✓		74.14	83.77	84.17	83.46	98.98
	✓		✓	74.86	84.27	83.24	85.50	99.03

Table 5. Ablation analysis on OSD dataset [37]. CAFT indicates the usage of CAFT blocks. HL and SL denote the use of hard labels and soft labels synthesized from our SAR-JointNet.

ing SAROSS-Net within the DAKD pipeline.

Next, in Tab. 5, we validate the effect of our CAFT block and DAKD pipeline, by utilizing the ResNet-34 [29] encoder with $21.8M$ parameters and the MambaVision-B [28] encoder with $97.7M$ parameters. As shown, simply incorporating CAFT blocks into skip connections improves performance for both encoders. Specifically, the mIoU increases by 1.78% and 1.39% for the ResNet-34 [29] and the MambaVision-B [28] encoders respectively, indicating that our CAFT blocks effectively handle encoded features from noisy SAR images, regardless of the encoder. Next, we compare the effect of the label type generated from our SAR-JointNet for segmentation. One-hot encoded segmentation masks (hard labels) improve the segmentation performance of both models due to the increased training data. The usage of logit values (soft labels) generated by our SAR-JointNet further improves the performance, as the logits carry richer information for transferring knowledge to student models. These results demonstrate that our DAKD pipeline effectively augments the dataset and enhances the generalizability of student models.

6. Conclusion

In this paper, a novel DAKD pipeline was introduced to enhance SAR image-based oil spill segmentation performance by addressing data scarcity and limited label availability. Our diffusion-based SAR-JointNet jointly generates image-label pairs efficiently with the introduction of a balancing factor, which ensures that information levels across modalities are effectively harmonized during the diffusion process. Additionally, we not only used image-label pairs from the SAR-JointNet for data augmentation but also *firstly* incorporated knowledge distillation. Logit-based labels generated by SAR-JointNet enable effective knowledge transfer to the segmentation model, SAROSS-Net. Lastly, we proposed the SAROSS-Net using Context-Aware Feature Transfer blocks, enabling robust segmentation even in SAR images with inherent speckle noises. Through this approach, we achieved state-of-the-art performance in SAR oil spill segmentation with *significantly large margins*.

DAKD: Data Augmentation and Knowledge Distillation using Diffusion Models for SAR Oil Spill Segmentation

Supplementary Material

In this supplementary material, we present a comprehensive analysis of our Data Augmentation and Knowledge Distillation (DAKD) pipeline for SAR oil spill segmentation. First, we provide a detailed derivation of the balancing factor b and describe the process of the training and inference of SAR-JointNet. We further present additional experimental results, showing the effectiveness of the proposed balancing factor b , CAFT blocks, and the impact of augmented dataset scale on segmentation performance. We also discuss a limitation of our DAKD pipeline. Lastly, we present additional qualitative results of the generated data from our SAR-JointNet, along with comparisons between our SAROSS-Net with DAKD pipeline and other SAR oil spill segmentation and natural image segmentation methods, such as CBD-Net [81], SegFormer [67], and DeepLabV3+ [11].

A. Detailed Derivation of the Balancing Factor

In Sec. 3.2.1, we introduce a balancing factor b to handle the retained information levels across modalities, specifically between SAR images and segmentation masks. To quantify the retained information level, we utilize signal-to-noise ratio (SNR), calculated via the 2D Discrete Fourier Transform (DFT). For image $\mathbf{x} \in \mathbb{R}^{H \times W}$, the 2D DFT is defined as:

$$F_{\mathbf{x}}(u, v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \mathbf{x}(h, w) \cdot e^{-i2\pi(\frac{uh}{H} + \frac{vw}{W})}, \quad (11)$$

where $F_{\mathbf{x}}(u, v)$ represents the complex frequency component of \mathbf{x} at the spatial frequency (u, v) . Each complex frequency component of a scaled image $a \cdot \mathbf{x}$ can be expressed as:

$$F_{a\mathbf{x}}(u, v) = a \cdot F_{\mathbf{x}}(u, v). \quad (12)$$

The mean power of image \mathbf{x} in frequency domain is defined as:

$$P(\mathbf{x}) = \mathbb{E}[F_{\mathbf{x}}(u, v) \cdot F_{\mathbf{x}}^*(u, v)], \quad (13)$$

where $*$ denotes the complex conjugate. For the scaled image $a \cdot \mathbf{x}$, its mean power is given as:

$$\begin{aligned} P(a\mathbf{x}) &= \mathbb{E}[F_{a\mathbf{x}}(u, v) \cdot F_{a\mathbf{x}}^*(u, v)] \\ &= \mathbb{E}[aF_{\mathbf{x}}(u, v) \cdot aF_{\mathbf{x}}^*(u, v)] \\ &= a^2 \cdot \mathbb{E}[F_{\mathbf{x}}(u, v) \cdot F_{\mathbf{x}}^*(u, v)] \\ &= a^2 \cdot P(\mathbf{x}). \end{aligned} \quad (14)$$

Using this property, we derive the balancing factor b . In the forward diffusion process, the noise-corrupted SAR image

\mathbf{x}_t and its corresponding segmentation mask \mathbf{y}_t at timestep t are expressed as follows:

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \mathbf{y}_t = \alpha_t \mathbf{y}_0 + \sigma_t \boldsymbol{\epsilon}. \quad (15)$$

where α_t and σ_t are scalars determined by the noise scheduler, representing the degree of signal and noise at timestep t , respectively. The SNR is defined as the ratio of the mean power of the signal to the mean power of the noise. So, the SNRs for \mathbf{x}_t and \mathbf{y}_t are given by:

$$\text{SNR}(\mathbf{x}_t) = \frac{\alpha_t^2 P(\mathbf{x}_0)}{\sigma_t^2 P(\boldsymbol{\epsilon})}, \quad \text{SNR}(\mathbf{y}_t) = \frac{\alpha_t^2 P(\mathbf{y}_0)}{\sigma_t^2 P(\boldsymbol{\epsilon})}. \quad (16)$$

To balance the levels of retained information for a SAR image and its corresponding mask, we scale the original segmentation mask \mathbf{y}_0 by the balancing factor b , resulting in a scaled noise-corrupted segmentation mask \mathbf{y}'_t , which is given by:

$$\mathbf{y}'_t = \alpha_t b \mathbf{y}_0 + \sigma_t \boldsymbol{\epsilon}. \quad (17)$$

The SNR for \mathbf{y}'_t is then given by:

$$\begin{aligned} \text{SNR}(\mathbf{y}'_t) &= \frac{P(\alpha_t b \mathbf{y}_0)}{P(\sigma_t \boldsymbol{\epsilon})} \\ &= \frac{\alpha_t^2 b^2 P(\mathbf{y}_0)}{\sigma_t^2 P(\boldsymbol{\epsilon})} \\ &= b^2 \cdot \text{SNR}(\mathbf{y}_t). \end{aligned} \quad (18)$$

The ratio of the retained information levels between \mathbf{x}_t and \mathbf{y}'_t is expressed as follows:

$$\frac{\text{SNR}(\mathbf{x}_t)}{\text{SNR}(\mathbf{y}'_t)} = \frac{\text{SNR}(\mathbf{x}_t)}{b^2 \cdot \text{SNR}(\mathbf{y}_t)}. \quad (19)$$

To make this ratio equal to 1, we obtain the balancing factor b as follows:

$$\begin{aligned} b &= \sqrt{\frac{\text{SNR}(\mathbf{x}_t)}{\text{SNR}(\mathbf{y}_t)}} \\ &= \sqrt{\frac{P(\mathbf{x}_0)}{P(\mathbf{y}_0)}}. \end{aligned} \quad (20)$$

This balancing factor ensures maintaining balanced information levels of different modalities between SAR images and masks in the forward diffusion process, which is critical for the successful training of SAR-JointNet. That is, this adjustment helps SAR-JointNet to focus equally on the features extracted from SAR images and segmentation masks, enabling more enhanced joint generation.

Algorithm 1 SAR-JointNet Training Algorithm

Input: $\mathbf{x}_0 \in \mathbb{R}^{H \times W \times 1}$, $\tilde{\mathbf{y}}_0 \in \{0, 1\}^{H \times W \times C}$
Output: loss

```

Function train_loss( $\mathbf{x}_0, \tilde{\mathbf{y}}_0$ ):
    # Normalize one-hot encoded segmentation mask
     $\mathbf{y}_0 \leftarrow \tilde{\mathbf{y}}_0 \cdot 2 - 1$ 

    # Scale with balancing factor  $b$ 
     $\mathbf{y}'_0 \leftarrow \mathbf{y}_0 \cdot b$ 

    # Add noise
     $t \sim \text{Uniform}(1, T)$ 
     $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ 
     $(\mathbf{x}_t, \mathbf{y}'_t) \leftarrow \alpha_t \cdot (\mathbf{x}_0, \mathbf{y}'_0) + \sigma_t \cdot \epsilon$ 

    # Predict  $\hat{\mathbf{x}}_\theta, \hat{\mathbf{y}}_\phi$ 
     $\hat{\mathbf{x}}_\theta, \hat{\mathbf{y}}_\phi \leftarrow \text{SAR-JointNet}((\mathbf{x}_t, \mathbf{y}'_t), t)$ 

    # Compute loss
     $\mathcal{L}_x \leftarrow \text{L2}(\hat{\mathbf{x}}_\theta, \mathbf{x}_0)$ 
     $\mathcal{L}_y \leftarrow \text{cross\_entropy}(\hat{\mathbf{y}}_\phi, \tilde{\mathbf{y}}_0)$ 

     $\mathcal{L}_{\text{joint}} \leftarrow \mathcal{L}_x + \mathcal{L}_y$ 
    return  $\mathcal{L}_{\text{joint}}$ 
```

B. Training and Inference of SAR-JointNet

B.1. Training of SAR-JointNet

Algorithm 1 summarizes the joint learning of our SAR-JointNet, specifically our training stages 2 and 3. As we described, an one-hot encoded segmentation mask $\tilde{\mathbf{y}}_0$ is normalized and scaled by the proposed balancing factor b . Then, we add noise to the scaled segmentation mask \mathbf{y}'_0 and a SAR image \mathbf{x}_0 with respect to a randomly sampled timestep t . Our SAR-JointNet jointly predicts the denoised output $\hat{\mathbf{x}}_\theta$ for SAR image reconstructions and logits $\hat{\mathbf{y}}_\phi$ for segmentation mask generations. For SAR image generation learning, we use an L2 loss between $\hat{\mathbf{x}}_\theta$ and \mathbf{x}_0 . For segmentation mask generation learning, we use a cross-entropy loss between $\hat{\mathbf{y}}_\phi$ and $\tilde{\mathbf{y}}_0$, following [14].

B.2. Inference of SAR-JointNet

Algorithm 2 illustrates the inference process of our SAR-JointNet. From the initial noise sampled from the Gaussian distribution, our SAR-JointNet predicts denoised output, $\hat{\mathbf{x}}_\theta$ and $\hat{\mathbf{y}}_\phi$. To match the range of the predicted logits for segmentation masks to that of $\mathbf{y}'_0 \in \{-b, b\}$, we rescale $\hat{\mathbf{y}}_\phi$ as follows:

$$\mathbf{y}_{\text{pred}} = (\text{softmax}(\hat{\mathbf{y}}_\phi) \cdot 2 - 1) \cdot b. \quad (21)$$

Using the step function of DDIM scheduler [58], we compute the previous sample \mathbf{x}_{t-1} and \mathbf{y}'_{t-1} using $\hat{\mathbf{x}}_\theta$, \mathbf{y}_{pred} , \mathbf{x}_t , \mathbf{y}'_t , and timestep t . After completing the reverse diffusion process, our SAR-JointNet outputs \mathbf{x}_0 that is a gener-

Algorithm 2 SAR-JointNet Inference Algorithm

```

# Initialize
 $\mathbf{x}_T, \mathbf{y}'_T \sim \mathcal{N}(0, \mathbf{I})$ 

for  $t = T$  to 1 do
    # Predict  $\hat{\mathbf{x}}_\theta, \hat{\mathbf{y}}_\phi$ 
     $\hat{\mathbf{x}}_\theta, \hat{\mathbf{y}}_\phi \leftarrow \text{SAR-JointNet}((\mathbf{x}_t, \mathbf{y}'_t), t)$ 

    # Rescale the predicted logit  $\hat{\mathbf{y}}_\phi$ 
     $\mathbf{y}_{\text{pred}} \leftarrow (\text{softmax}(\hat{\mathbf{y}}_\phi) \cdot 2 - 1) \cdot b$ 

    # Compute deterministic next step
     $\mathbf{x}_{t-1}, \mathbf{y}'_{t-1} \leftarrow \text{ddim\_step}(\hat{\mathbf{x}}_\theta, \mathbf{y}_{\text{pred}}), (\mathbf{x}_t, \mathbf{y}'_t), t)$ 
end for

# Return denoised SAR image and logits
return  $\mathbf{x}_0, \hat{\mathbf{y}}_\phi$ 
```

b	FID (\downarrow)	IS (\uparrow)	mIoU (%)	F1 (%)
0.3	27.84	2.196	74.18	83.78
0.528	30.64	2.186	74.86	84.27
0.7	35.53	2.158	74.42	83.98
1	44.81	2.158	74.08	83.76

Table 6. Ablation analysis on the balancing factor b . We assess the quality of generated SAR images and the segmentation performance of our SAROSS-Net on the OSD dataset [37].

ated SAR image, and $\hat{\mathbf{y}}_\phi$ that serves as a knowledge distillation signal for training SAROSS-Net.

C. Effect of the Balancing Factor

In this section, we verify the effectiveness of the balancing factor b derived from the ratio of SNR between the SAR image and its corresponding segmentation mask in SAR-JointNet. We compare the quality of generated data from SAR-JointNet trained using $b = 0.528$, derived from Eq. (20) and using other values 0.3, 0.7, and 1 on the OSD dataset [37].

As shown in Fig. 6, noise-corrupted segmentation masks without applying b retain significantly more information than noise-corrupted SAR images at the same timestep t . As discussed earlier, smaller values of b reduce the retained information in the noise-corrupted segmentation masks. This encourages SAR-Net within SAR-JointNet to generate SAR images without heavily relying on the mask features transferred through the interconnection layers between SAR-Net and Mask-Net. Consequently, smaller b values lead to improved SAR image quality, as evidenced by better FID and IS in Tab. 6. However, excessively small b values overly suppress the information in the segmentation masks, resulting in information levels even lower than those of the SAR images. This imbalance makes it difficult for Mask-Net to learn effectively.

Using the predefined value $b = 0.3$, smaller than the

Methods	Params. (M)	FLOPs (G)	Time (ms)	Memory (MB)	mIoU (%)
No module	-	-	39.52	-	71.99
3x3 Conv	+6.95	+37.06	48.36	+37	70.95
CAFT	+0.89	+11.45	45.28	+11	73.38

Table 7. Ablation analysis on the usage of CAFT blocks in skip connection of SAROSS-Net on the OSD dataset [37].

derived value of $b = 0.528$, has improved the quality of generated SAR images in terms of FID and IS, but has led to lower segmentation performance in terms of mIoU and F1 scores of 74.18% and 83.78%, respectively on the OSD dataset [37]. The balancing factor b with the value of 0.528, derived from Eq. (20), demonstrates strong generation quality and achieves the best segmentation performance with 74.86% and 84.27% in mIoU and F1 scores, respectively, for oil spill segmentation on OSD dataset [37].

D. Effect of CAFT Blocks

In this section, we conduct an ablation study to validate the performance and effectiveness of our proposed Context-Aware Feature Transfer (CAFT) blocks in skip connection layers. Specifically, we compare our model incorporating CAFT blocks against two baselines with: (i) no additional module in each skip connection, and (ii) 3×3 convolution blocks that replace all CAFT blocks in the skip connections of Fig. 3. For all three models, we leverage the pretrained MambaVision-B [28] as the encoder backbone. To evaluate effectiveness, we compare the number of parameters, FLOPs, inference time, and memory allocation, which are measured using a single 512×512 image input (batch size of 1), while the inference time is calculated as the average over 300 runs. All evaluations are conducted on an NVIDIA RTX 2080 GPU. Additionally, we assess segmentation performance using the mIoU metric on the OSD dataset [37].

As shown in Tab. 7, our full model with CAFT blocks (SAROSS-Net) achieves a higher mIoU while maintaining a lower parameter count, reduced FLOPs, shorter inference time, and lower memory consumption, compared to the 3×3 convolution-based baseline. This highlights the efficiency and superior performance of our proposed CAFT blocks in enhancing the capability to selectively transfer high-frequency features to the decoder. As demonstrated in Fig. 7, our SAROSS-Net performs more precise segmentation, which highlights the capability to selectively transfer high-frequency features from noisy SAR images. It effectively captures and refines crucial information, significantly improving segmentation performance.

Aug. Data	mIoU (%)	F1 (%)	R (%)	P (%)	A (%)
0	73.38	83.25	81.93	84.70	98.82
1×	73.81	83.54	84.35	82.90	98.79
2×	74.65	84.13	81.85	86.75	99.00
4×	74.86	84.27	83.24	85.50	99.03

Table 8. Segmentation performance on the impact of the amount of the augmented dataset for training of oil spill segmentation on the OSD dataset [37].

E. Impact of the Amount of Augmented Dataset

We further evaluate the impact of the amount of the augmented dataset from SAR-JointNet on segmentation performance. By systematically varying the amount of the augmented samples for the training of SAROSS-Net, we aim to analyze how additional *generated* data contributes to the model’s learning ability to generalize and accurately segment target regions. Tab. 8 shows the segmentation performance on the usage amounts of augmented samples generated by SAR-JointNet. Note that the original OSD dataset [37] contains 1,002 samples, while the numbers of generated samples are 1,120 (1×), 2,240 (2×), and 4,480 (4×), corresponding to the number of original training samples. Both the original training dataset and generated augmented datasets are used together to train SAROSS-Net. As demonstrated, the segmentation performance consistently improves as the size (number) of the augmented samples increases, highlighting the importance of *generative data augmentation* in enhancing the segmentation performance.

F. Limitations

While our DAKD pipeline successfully generates SAR images and their corresponding segmentation masks using our SAR-JointNet, it is currently constrained to a resolution of 256×256 . Extending SAR-JointNet, a pixel-domain diffusion model, to generate spatially larger images, such as 512×512 , requires significantly more computational resources. Although the latent diffusion model (LDM) [52] has demonstrated successful generation capability for high-resolution natural images, we observed that LDM struggles to synthesize speckle noise of SAR images in the latent domain, which is typically reduced to 1/8 or 1/4 of the original image size. To address this limitation, we adopt a technique inspired by mosaic augmentation used in YOLOv4 [4], where four randomly selected 256×256 generated images and combine them to form a 512×512 composite image. While this method is practical, it does not fully resolve the challenge of high-resolution SAR image synthesis. Future work should aim to develop efficient models capable of generating high-resolution SAR images while preserving their unique noise characteristics.

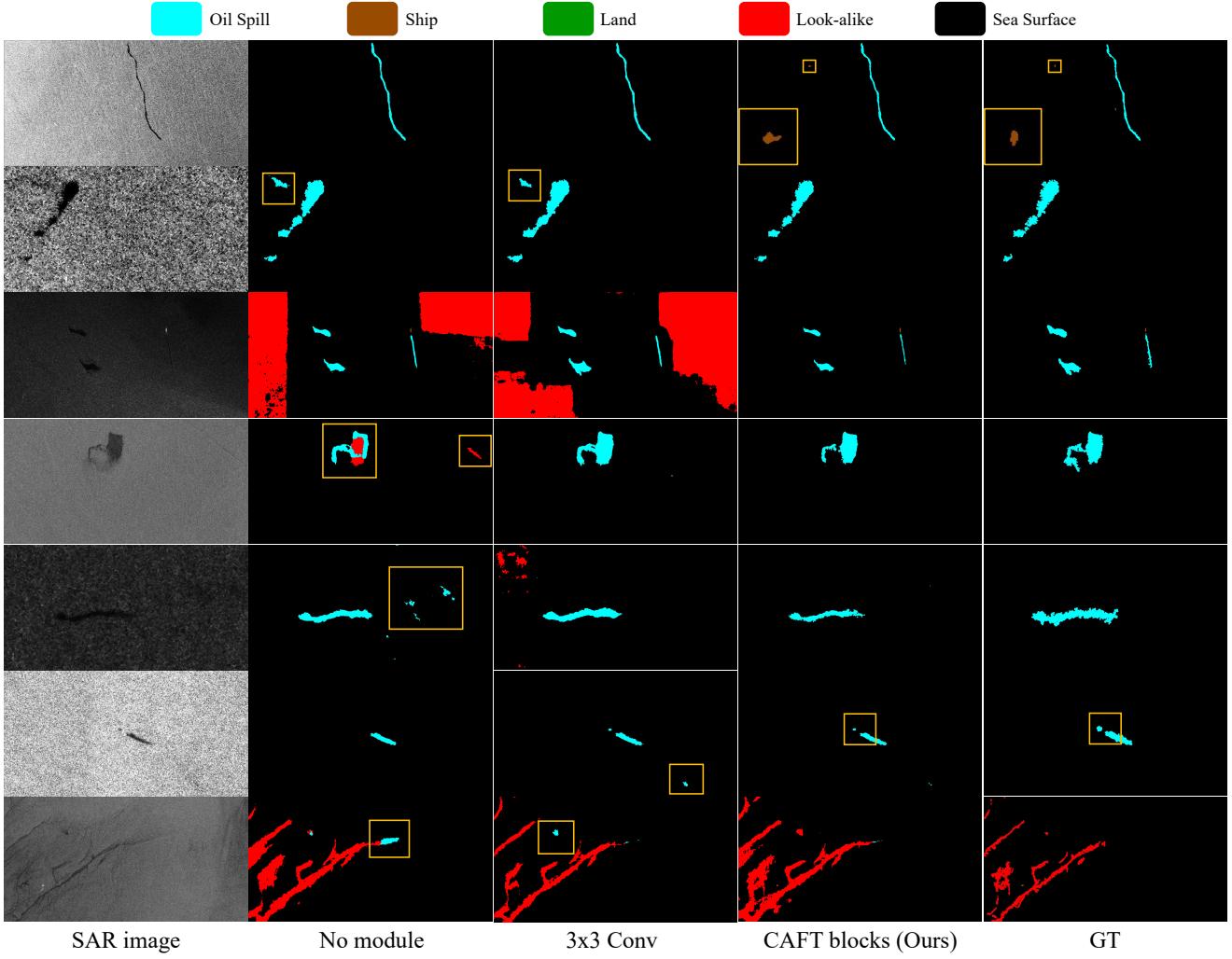


Figure 7. Qualitative comparison of segmentation results on SAR images for the OSD dataset [37] to see the effectiveness of CAFT blocks. The results demonstrate that our full model with CAFT blocks (SAROSS-Net) effectively captures high-frequency features based on the semantic features, producing more accurate and visually consistent segmentation, particularly in noisy SAR images.

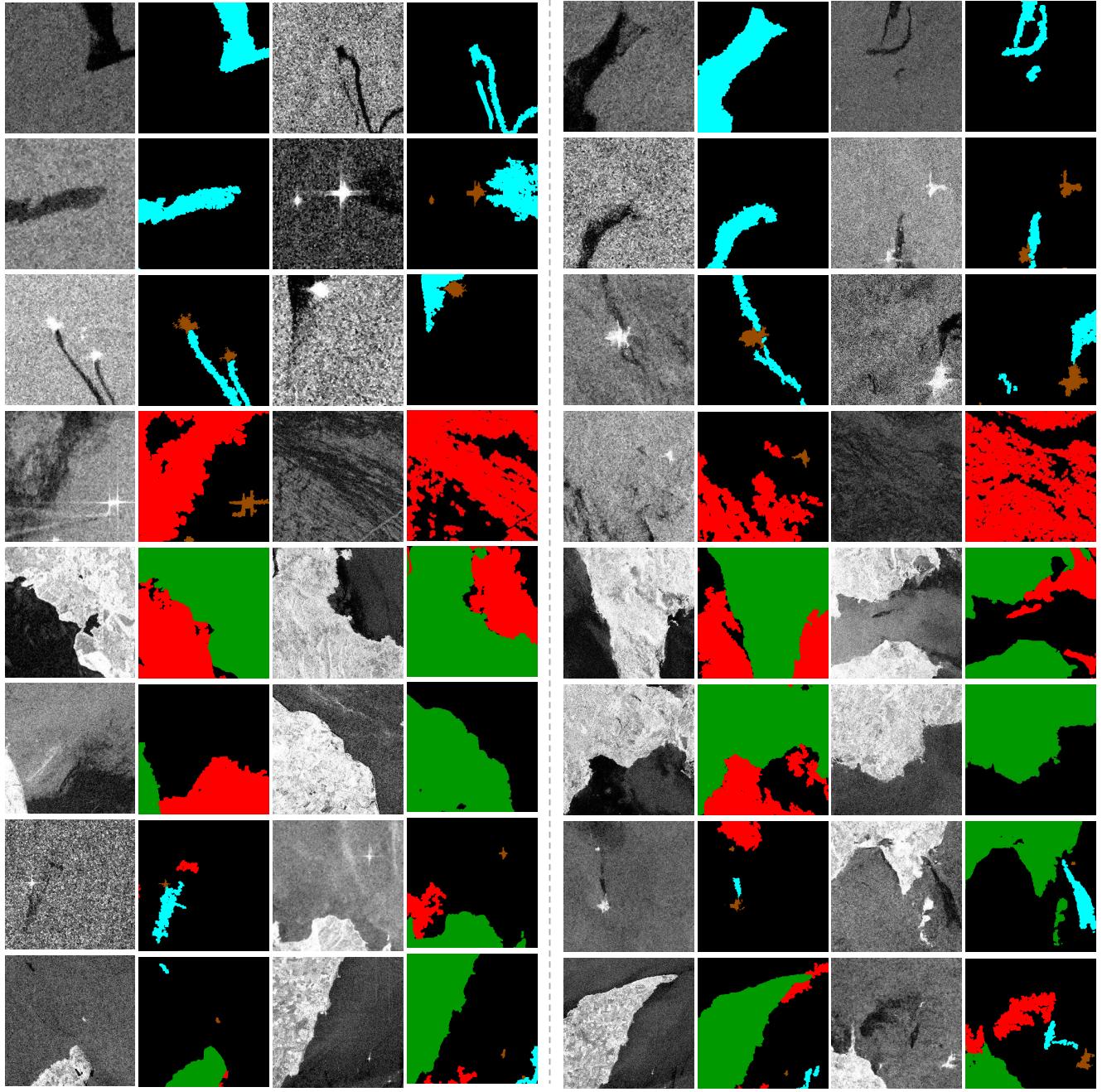
G. Additional Experimental Results

G.1. Data Generation Results from SAR-JointNet

We provide additional generated pairs of SAR images and their segmentation masks from our SAR-JointNet, comparing to the original training dataset (OSD dataset [37]) in Fig. 8. Furthermore, Fig. 9 and Fig. 10 show SAR images and segmentation masks from the original training dataset of SOS-Sentinel-1A and SOS-ALOS [81], as well as the respective pairs generated by SAR-JointNet. Our results demonstrate that SAR-JointNet generates realistic SAR images and segmentation masks with high correspondences, effectively capturing the characteristics of the original datasets.

G.2. Segmentation Results

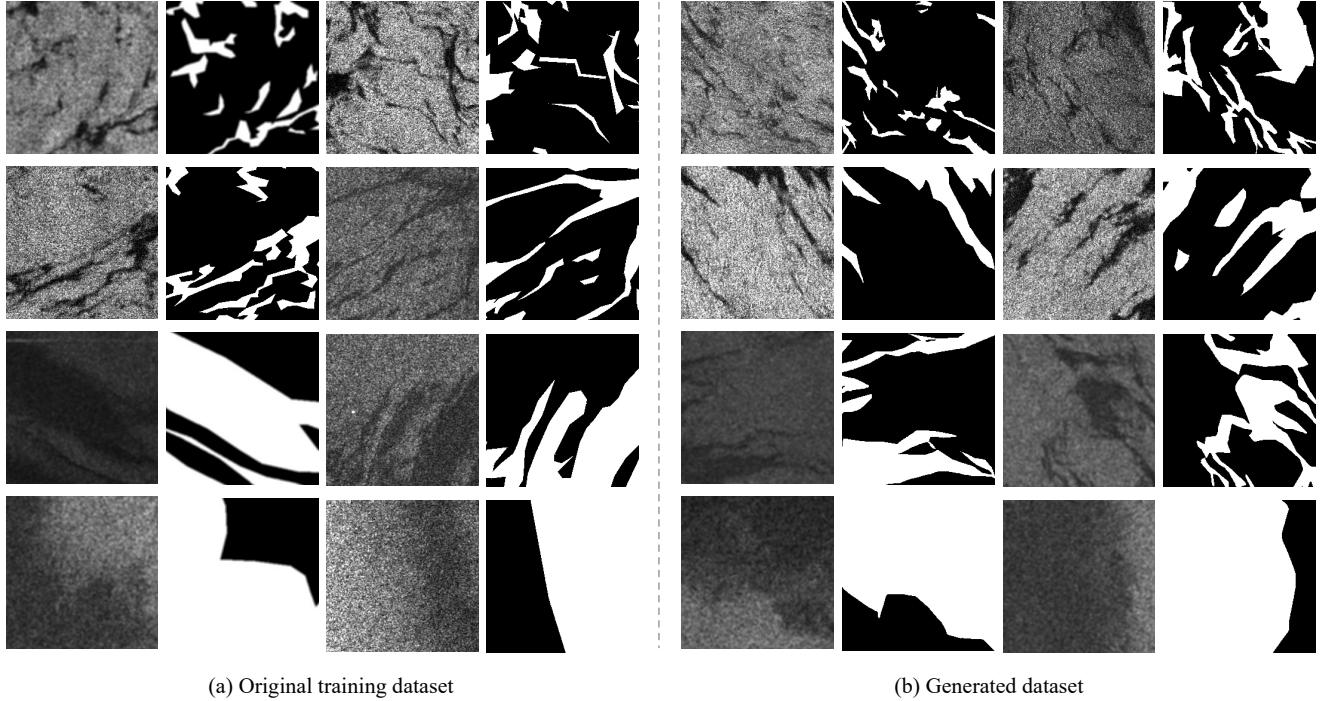
Fig. 11, Fig. 12, and Fig. 13 present qualitative results on the OSD dataset [37], SOS-ALOS, and SOS-Sentinel-1A dataset [81], respectively. The compared models include CBD-Net [81], SegFormer [67], and DeepLabV3+ [11]. Our proposed SAROSS-Net outperforms these models, achieving the highest performance on the noisy SOS dataset [81]. Notably, SAROSS-Net demonstrates a superior ability to identify oil spill regions, effectively handling the challenges posed by noisy data and selectively capturing high-frequency features crucial for accurate segmentation. These results emphasize the robustness and effectiveness of SAROSS-Net in addressing oil spill detection tasks.



(a) Original training dataset

(b) Generated dataset

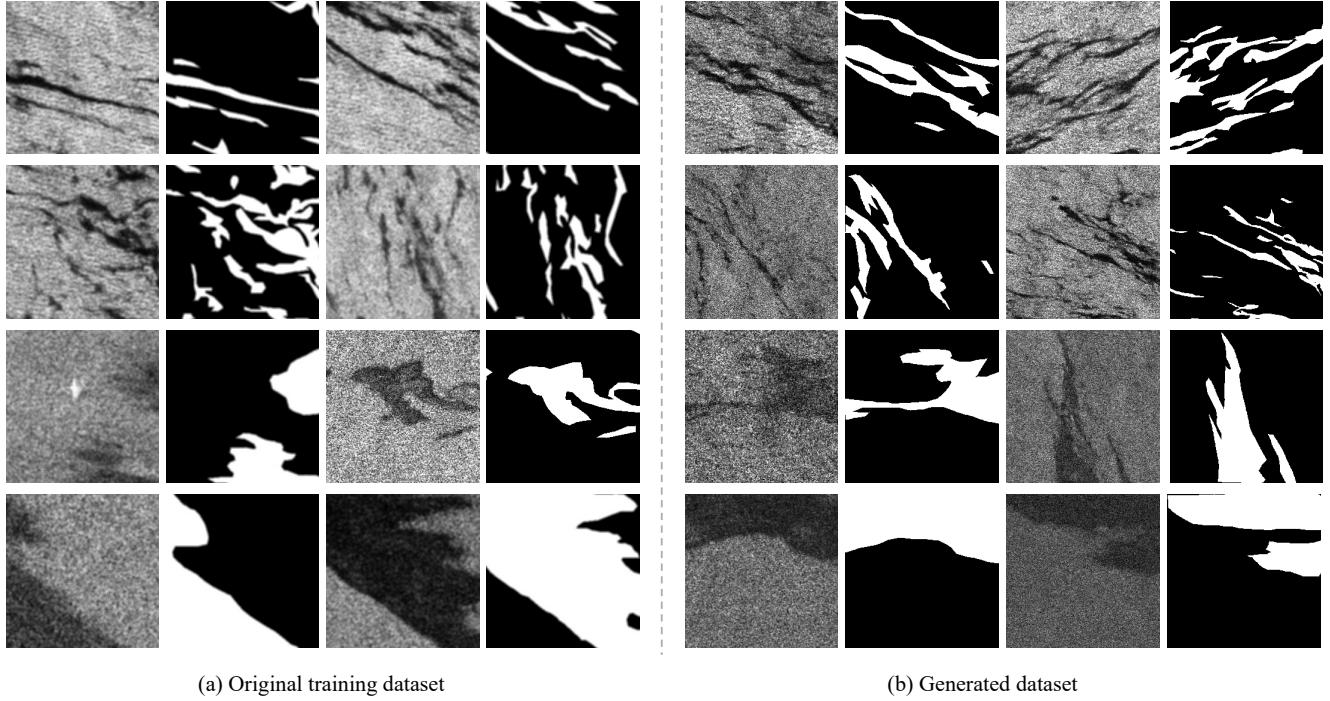
Figure 8. Qualitative results of generated SAR images and their corresponding generated segmentation masks. (a) shows several original training samples (SAR images and their associated segmentation masks) of the OSD dataset [37], (b) shows several generated samples from our SAR-JointNet.



(a) Original training dataset

(b) Generated dataset

Figure 9. Qualitative results of generated SAR images and their corresponding generated segmentation masks. (a) shows several original training samples (SAR images and their associated segmentation masks) of SOS-Sentinel-1A dataset [81], (b) shows several generated samples from our SAR-JointNet.



(a) Original training dataset

(b) Generated dataset

Figure 10. Qualitative results of generated SAR images and their corresponding generated segmentation masks. (a) shows several original training samples (SAR images and their associated segmentation masks) of SOS-ALOS dataset [81], (b) shows several generated samples from our SAR-JointNet.

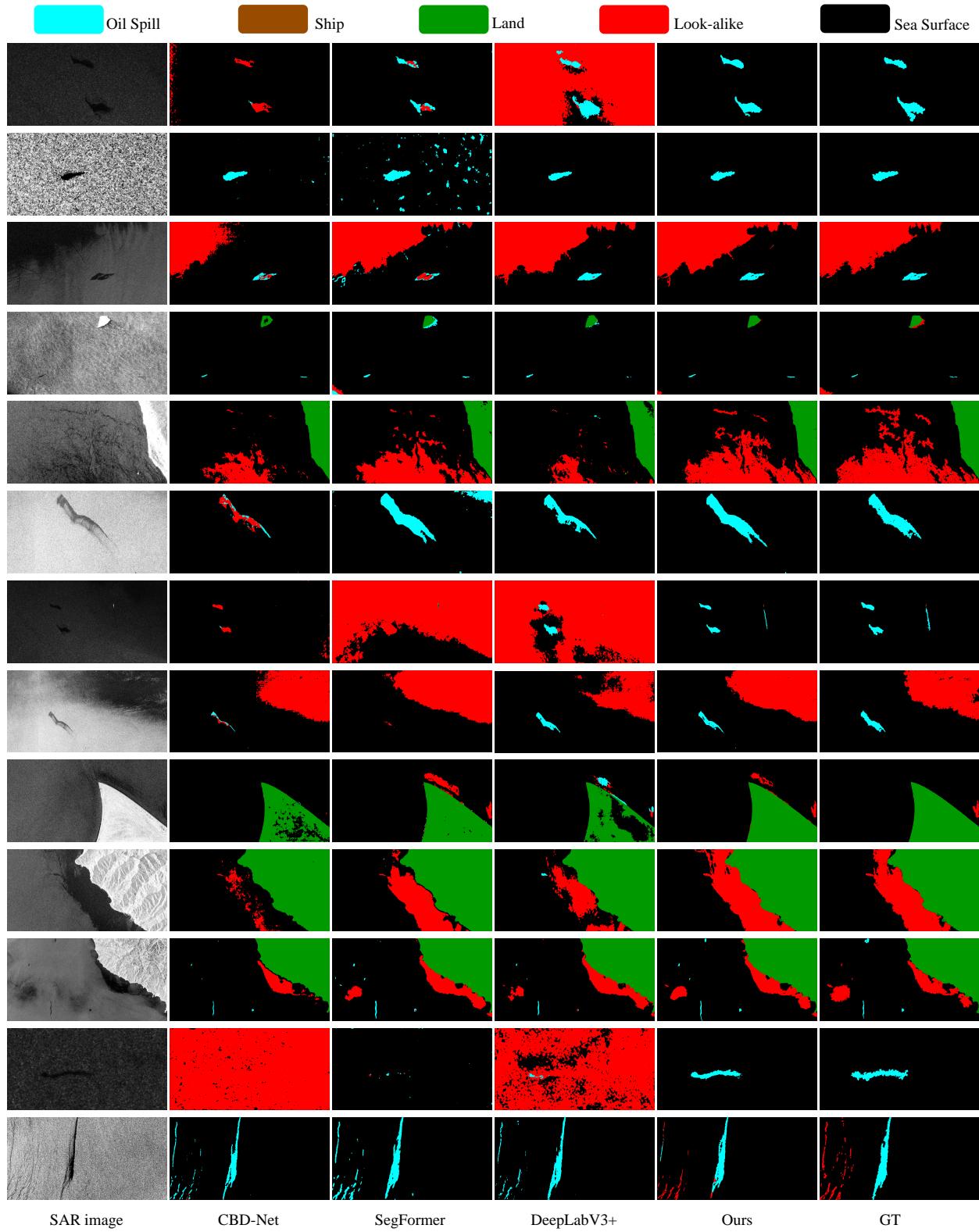


Figure 11. Qualitative comparison of segmentation results on the OSD dataset [37]. Each row shows a SAR input, four segmentation results by CBD-Net [81], SegFormer [67], DeepLabV3+ [11] and our SAROSS-Net, and a ground truth segmentation mask, respectively.

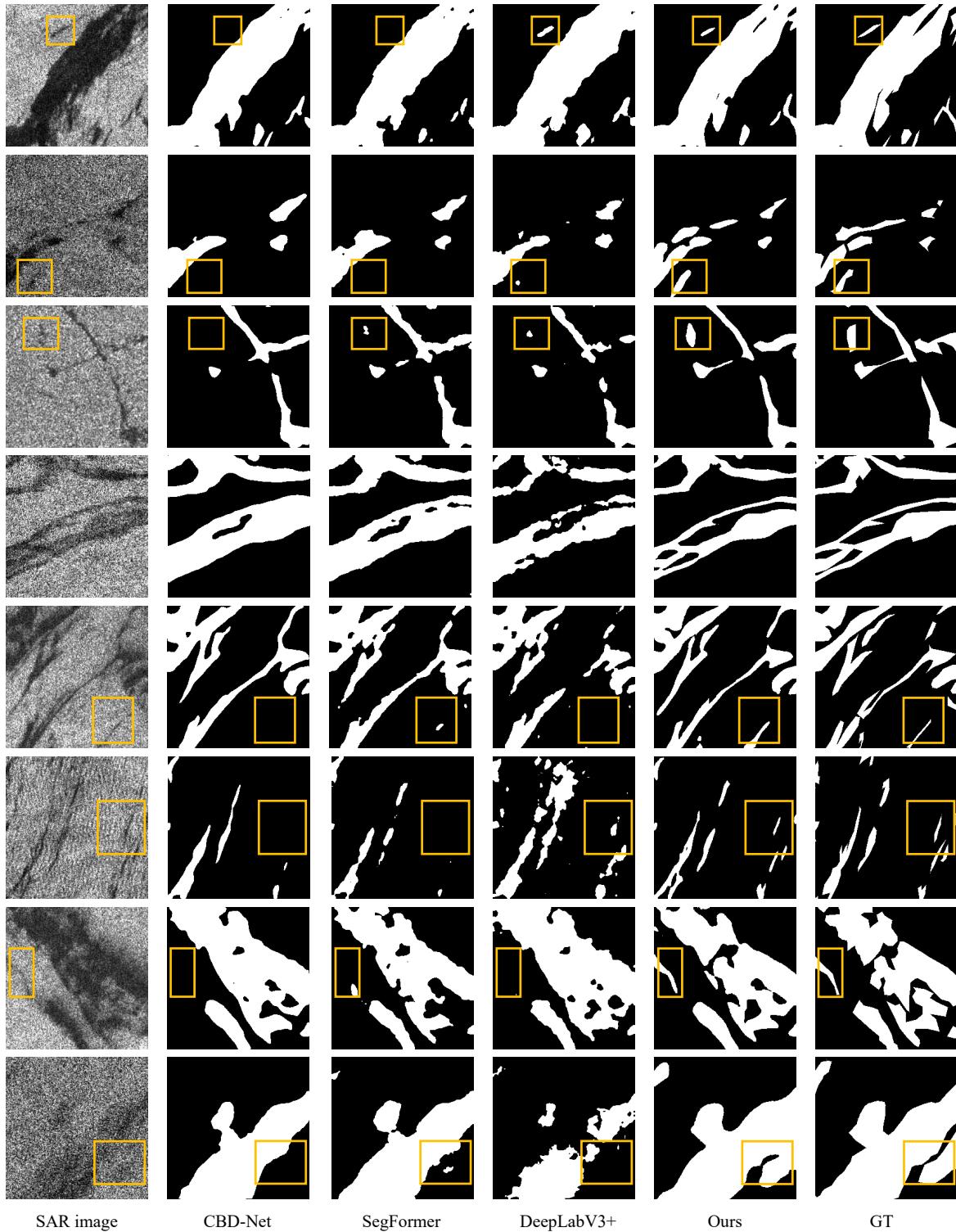


Figure 12. Qualitative comparison of segmentation results on SOS-ALOS dataset [81]. Each row shows a SAR input, four segmentation results by CBD-Net [81], SegFormer [67], DeepLabV3+ [11] and our SAROSS-Net, and a ground truth segmentation mask, respectively.

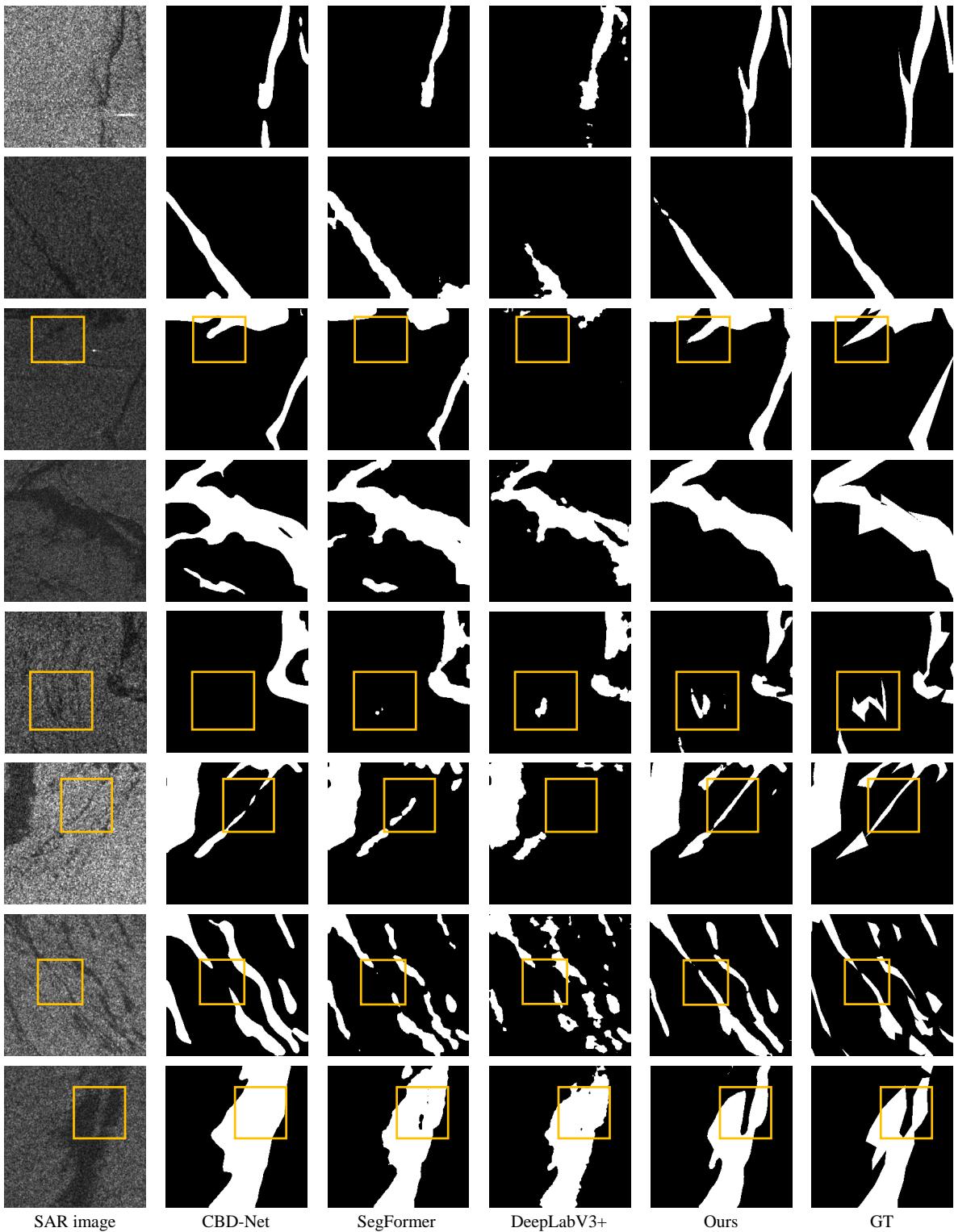


Figure 13. Qualitative comparison of segmentation results on SOS-Sentinel-1A dataset [81]. Each row shows a SAR input, four segmentation results by CBD-Net [81], SegFormer [67], DeepLabV3+ [11] and our SAROSS-Net, and a ground truth segmentation mask, respectively.

References

- [1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2, 3
- [2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2
- [3] F. M. Bianchi, M. M. Espeseth, and N. Borch. Large-scale detection and categorization of oil spills from sar images with deep learning. *Remote Sensing*, 12(14):2260, 2020. 2
- [4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 6, 11
- [5] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017. 7
- [6] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3
- [7] Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Geodiffusion: Text-prompted geometric control for object detection data generation. *arXiv preprint arXiv:2306.04607*, 2023. 3
- [8] Liang-Chieh Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7
- [9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 2016. 2
- [10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 7
- [11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2, 7, 9, 12, 15, 16, 17
- [12] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023. 4
- [13] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022. 4
- [14] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 909–919, 2023. 4, 10
- [15] Y. Chen, Y. Li, and J. Wang. An end-to-end oil-spill monitoring method for multisensory satellite images based on deep semantic segmentation. *Sensors*, 20(3):725, 2020. 2
- [16] Hyunho Choi and Jechang Jeong. Speckle noise reduction technique for sar images using statistical characteristics of speckle noise and discrete wavelet transform. *Remote Sensing*, 11(10), 2019. 1
- [17] Dario Cioni, Lorenzo Berlincioni, Federico Becattini, and Alberto Del Bimbo. Diffusion based augmentation for captioning and retrieval in cultural heritage. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, page 1699–1708. IEEE, 2023. 3
- [18] MMYOLO Contributors. MMYOLO: OpenMMLab YOLO series toolbox and benchmark. <https://github.com/open-mmlab/mmyolo>, 2022. 2
- [19] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. Multi-head attention: Collaborate instead of concatenate. *arXiv preprint arXiv:2006.16362*, 2020. 2
- [20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2
- [21] Martine Espeseth, Cathleen Jones, Benjamin Holt, Camilla Brekke, and Stine Skrunes. Oil-spill-response-oriented information products derived from a rapid-repeat time series of sar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:1–1, 2020. 1
- [22] Gongfan Fang, Kanya Mo, Xinchao Wang, Jie Song, Shitao Bei, Haofei Zhang, and Mingli Song. Up to 100× faster data-free knowledge distillation. *arXiv preprint arXiv:2112.06253*, 2022. 3
- [23] Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1257–1266, 2024. 2, 3
- [24] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 2
- [25] Yunxiang Fu, Chaoqi Chen, Yu Qiao, and Yizhou Yu. Dreamda: Generative data augmentation with diffusion models. *arXiv preprint arXiv:2403.12803*, 2024. 3
- [26] L. Gemme and S. G. Dellepiane. An automatic data-driven method for sar image segmentation in sea surface analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 56(5):2633–2646, 2018. 1
- [27] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2
- [28] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv preprint arXiv:2407.08083*, 2024. 2, 4, 6, 8, 11

- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8
- [30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3, 5
- [31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 6, 7
- [32] Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge diffusion for distillation. In *Advances in Neural Information Processing Systems*, pages 65299–65316. Curran Associates, Inc., 2023. 3
- [33] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2
- [35] M. Konik and K. Bradtke. Object-oriented approach to oil spill detection using envisat asar images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 118:37–52, 2016. 1
- [36] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2024. 2
- [37] M. Krestenitis, G. Orfanidis, K. Ioannidis, K. Avgerinakis, S. Vrochidis, and I. Kompatsiaris. Oil spill identification from satellite images using deep neural networks. *Remote Sensing*, 11(15):1762, 2019. 2, 5, 6, 7, 8, 10, 11, 12, 13, 15
- [38] I. K. Lee, A. Shamsoddini, X. Li, J. C. Trinder, and Z. Li. Extracting hurricane eye morphology from spaceborne sar images using morphological analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:115–125, 2016. 1
- [39] Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, and Zhigeng Pan. Online knowledge distillation for efficient pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11740–11750, 2021. 3
- [40] Zheng Li, Yuxuan Li, Penghai Zhao, Renjie Song, Xiang Li, and Jian Yang. Is synthetic data from diffusion models ready for knowledge distillation? *arXiv preprint arXiv:2305.12954*, 2023. 2, 3
- [41] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2
- [42] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 3
- [43] Shujun Liu, Guoqing Wu, Xinzhen Zhang, Kui Zhang, Pin Wang, and Yongming Li. Sar despeckling via classification-based nonlocal and local sparse representation. *Neurocomputing*, 219:174–185, 2017. 1
- [44] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2604–2613, 2019. 3
- [45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2
- [46] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017. 3
- [47] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [48] M. Nieto-Hidalgo, A. J. Gallego, P. Gil, and A. Pertusa. Two-stage convolutional neural network for ship and spill detection using sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9):5217–5230, 2018. 1
- [49] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019. 3
- [50] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3
- [51] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 2
- [52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 11
- [53] 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, pages 234–241. Springer, 2015. 7
- [54] A.-B. Salberg, O. Rudjord, and A. H. S. Solberg. Oil spill detection in hybrid-polarimetric sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 52(10):6521–6533, 2014. 1
- [55] Jacob Schnell, Jieke Wang, Lu Qi, Vincent Tao Hu, and Meng Tang. Generative data augmentation improves scribble-supervised semantic segmentation. In *CVPR 2024 Workshop SyntaGen: Harnessing Generative Models for Synthetic Visual Datasets*, 2024. 3
- [56] Y. Shu, J. Li, H. Yousif, and G. Gomes. Dark-spot detection from sar intensity imagery with spatial density thresholding

- for oil-spill monitoring. *Remote Sensing of Environment*, 114(9):2026–2035, 2010. 1
- [57] Prabhishhek Singh and Raj Shree. Analysis and effects of speckle noise in sar images. In *2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA) (Fall)*, pages 1–5, 2016. 1
- [58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6, 10
- [59] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*, page 240–248. Springer International Publishing, 2017. 5
- [60] Aysim Toker, Marvin Eisenberger, Daniel Cremers, and Laura Leal-Taixé. Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27695–27705, 2024. 3, 6
- [61] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 3
- [62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 2, 5
- [63] M. Vespe and H. Greidanus. Sar image quality assessment and indicators for vessel and oil spill detection. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4726–4734, 2012. 1
- [64] W. Wang, H. Sheng, S. Liu, Y. Chen, J. Wan, and J. Mao. An edge-preserving active contour model with bilateral filter based on hyperspectral image spectral information for oil spill segmentation. In *Proceedings of the Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–5, 2019. 1, 2
- [65] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 346–362. Springer, 2020. 3
- [66] Wenhui Wu, Man Sing Wong, Xinyu Yu, Guoqiang Shi, Coco Yin Tung Kwok, and Kang Zou. Compositional oil spill detection based on object detector and adapted segment anything model from sar images. *IEEE Geoscience and Remote Sensing Letters*, 2024. 2, 7
- [67] Enze Xie, Wenhui 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*, 34: 12077–12090, 2021. 7, 9, 12, 15, 16, 17
- [68] Jiahao Xie, Wei Li, Xiangtai Li, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. *International Journal of Computer Vision*, pages 1–20, 2024. 2
- [69] Xinyi Yu, Guanbin Li, Wei Lou, Siqi Liu, Xiang Wan, Yan Chen, and Haofeng Li. Diffusion-based data augmentation for nuclei image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 592–602. Springer, 2023. 2, 3
- [70] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 5
- [71] Feng Zhang, Xiatian Zhu, and Mao Ye. Fast human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3517–3526, 2019. 3
- [72] Jingyang Zhang, Shiwei Li, Yuanxun Lu, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, and Yao Yao. Jointnet: Extending text-to-image diffusion for dense distribution modeling. *International Conference on Learning Representations (ICLR)*, 2024. 3, 6, 7
- [73] Manlin Zhang, Jie Wu, Yuxi Ren, Ming Li, Jie Qin, Xuefeng Xiao, Wei Liu, Rui Wang, Min Zheng, and Andy J Ma. Diffusionengine: Diffusion model is scalable data engine for object detection. *arXiv preprint arXiv:2309.03893*, 2023. 3
- [74] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11953–11962, 2022. 3
- [75] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2, 7
- [76] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 4
- [77] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 2
- [78] Qiqi Zhu, Zhen Li, Yanan Zhang, and Qingfeng Guan. Building extraction from high spatial resolution remote sensing images via multiscale-aware and segmentation-prior conditional random fields. *Remote Sensing*, 12(23):3983, 2020. 2
- [79] Qiqi Zhu, Weihuan Deng, Zhuo Zheng, Yanfei Zhong, Qingfeng Guan, Weihua Lin, Liangpei Zhang, and Deren Li. A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification. *IEEE Transactions on Cybernetics*, 52(11):11709–11723, 2021.
- [80] Qiqi Zhu, Yanan Zhang, Lizeng Wang, Yanfei Zhong, Qingfeng Guan, Xiaoyan Lu, Liangpei Zhang, and Deren Li.

A global context-aware and batch-independent network for road extraction from vhr satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:353–365, 2021.

[2](#)

- [81] Qiqi Zhu, Yanan Zhang, Ziqi Li, Xiaorui Yan, Qingfeng Guan, Yanfei Zhong, Liangpei Zhang, and Deren Li. Oil spill contextual and boundary-supervised detection network based on marine sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–10, 2022. [2](#), [5](#), [6](#), [7](#), [9](#), [12](#), [14](#), [15](#), [16](#), [17](#)