LAST: Utilizing Synthetic Image Style Transfer to Tackle Domain Shift in Aerial Image Segmentation

Yubo Wang, Ruijia Wen, Hiroyuki Ishii and Jun Ohya
Department of Modern Mechanical and Engineering, Waseda University, Tokyo, Japan
{*bobwang, wenruijia*}*@toki.waseda.jp,* {*hiro.ishii, ohya*}*@waseda.jp*

Keywords: Domain Shift, Style Transfer, Aerial Image Processing, Semantic Segmentation.

Abstract: Recent deep learning models often struggle with performance degradation due to domain shifts. Addressing domain adaptation in aerial image segmentation is challenging due to the limited availability of training data. To tackle this, we utilized the Unreal Engine to construct a synthetic dataset featuring images captured under diverse conditions such as fog, snow, and nighttime settings. We then proposed a latent space style transfer model that generates alternate domain versions based on the real aerial dataset. This approach eliminates the need for additional annotations on shifted domain data. We benchmarked nine different *state-of-the-art* segmentation methods on the ISPRS Vaihingen, Potsdam datasets, and their shifted foggy domains. Extensive experiments reveal that domain shift leads to significant performance drops, with an average decrease of **-3.46% mIoU** on Vaihingen and **-5.22% mIoU** on Potsdam. Finally, we adapted the model to perform well in the shifted domain, achieving improvements of **+2.97% mIoU** on Vaihingen and **+3.97% mIoU** on Potsdam, while maintaining its effectiveness in the original domain.

1 INTRODUCTION

1.1 Domain Shift Caused Model Degradation in Aerial Image

Aerial Image Segmentation (AIS) is an essential task for various city monitoring purposes, such as environmental surveillance, target localization, and disaster response (Pi et al., 2020; Wang et al., 2022; Liang et al., 2023). With semantic segmentation models trained on large-scale annotated data, humans can easily extract abundant geo-spatial information from aerial images captured by drones or satellites (Li et al., 2021; Wang et al., 2024a; Toker et al., 2024).

However, while the performance of semantic segmentation algorithms has surged on common benchmarks, progress in handling the domain shift of unseen environmental conditions is still stagnant (Dai and Van Gool, 2018; Michaelis et al., 2019; Sun et al., 2022). We demonstrate that the aerial segmentation performance of algorithms is prone to significant degradation due to **Domain Shift**, i.e., the transfer from one domain to another. In Figure 1, we illustrate this phenomenon by comparing the original data in the ISPRS datasets (Gerke, 2012; Rottensteiner et al., 2014) with our generated domain-shifted versions.

Figure 2 illustrates that even within the same scene, changing weather conditions and varying lighting levels pose challenges for aerial image segmentation algorithms. Specifically, we evaluated nine state-of-the-art segmentation models on the ISPRS dataset and its domain-shifted version. The results show that after transferring the data from its original, intact domain to a shifted fog domain, there is an average mIoU deterioration of **-3.46%** on the Vaihingen dataset (398 RGB images at 512×512 resolution) and **-5.22%** on the Potsdam dataset (2016 RGB images at 512×512 resolution). Notably, compared to the original intact data, the illumination in the shifted fog images is significantly reduced, and the weather conditions have changed from clear skies to foggy, representing a typical domain shift. However, the image content, layout, and geo-spatial information between the original and foggy data remain unchanged.

Closing the gap between model performance in the original domain and the shifted domain is a valuable problem to address. An intuitive solution is to incorporate multi-domain data into the model training process. The performance of aerial image segmentors significantly relies on the availability of training data. Although data from adverse domains is essential to improve the robustness of aerial image segmentation models, such data—including aerial images

Wang, Y., Wen, R., Ishii, H. and Ohya, J.
LAST: Utilizing Synthetic Image Style Transfer to Tackle Domain Shift in Aerial Image Segmentation.
DOI: 10.5220/0013145000003905

In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 32-42
ISBN: 978-989-758-730-6; ISSN: 2184-4313

Figure 1: **Examples of domain shift in aerial image** , in which the up row are the original aerial image from ISPRS Vaihingen and Potsdam (Gerke, 2012; Rottensteiner et al., 2014), while the bottom are the corresponding domain shifted imagery in Foggy condition. Notably, the image information including scene, target remained the same, on the contrary, the weather and illumination changed.

captured under low illumination and harsh weather conditions—are lacking in the current aerial image benchmarks (Waqas Zamir et al., 2019; Gerke, 2012; Rottensteiner et al., 2014).

1.2 Recent Development on Image Generation and Synthesis

Recently, significant triumph has been achieved by generative model, which aims to mimic human's ability on yielding various modalities, such as GPT-series (Brown, 2020) in Natural Language Processing and stable-diffusion (Rombach et al., 2022) in Computer Vision. Prior methodologies like Generative adversarial network (GAN)-based methods (Goodfellow et al., 2014; Zhu et al., 2017; Zhang et al., 2017; Brock, 2018; Karras et al., 2019; Zhang et al., 2019) and Variational autoencoder (VAE)-based methods (Kingma, 2013; Vahdat and Kautz, 2020) demonstrate remarkable performance in yielding realistic samples. Despite the sccess, training instability is a well-known issue, as GANs require a delicate balance between the generator and discriminator, which can lead to problems like mode collapse—where the generator produces limited diversity in outputs.

In addition, instead of traditional diffusion models (DMs) that denoise the input x in image-scale (Sohl-Dickstein et al., 2015; Ho et al., 2020), current Latent diffusion model (LDMs) (Ramesh et al., 2022; Rombach et al., 2022; Zhang et al., 2023; Luo et al., 2023) adopt a VAE-like Encoder $\mathcal{E}$ and Decoder $\mathcal{D}$ structure. LDMs first compress the input into a latent representation $z = \mathcal{E}(\mathrm{x})$, afterwards deploy diffusion process within latent space, such that decoder outputs $\tilde{x}$ is the reconstructed input x. With the hallmark of achieving a favorable trade-off between reducing computational and memory costs and maintaining high resolution and quality synthesis, operating on smaller spatial latent representations of the input has become a popular framework for recent generative models (Li et al., 2023; Khanna et al., 2023; Peebles and Xie, 2023), i.e., LDMs. However, the tedious sampling step of the diffusion model renders it highly inefficient for use with large-scale datasets (Song et al., 2020; Wang et al., 2024b; Karras et al., 2024; Gong et al., 2024).

Image Style Transfer. (Gatys et al., 2016; Deng et al., 2022; Brooks et al., 2023; Wang et al., 2023; Sohn et al., 2024; Chung et al., 2024) is a practical generative task that aims to extract the style texture information of one reference image and merge it with the content from another semantic image. The prior methods can synthesise vivid and diverse results, such as converting a landscape photo into a painterly oil artwork or creating a cartoon version of a person's portrait. However, for *de-facto* domain shifts in aerial imagery, performance aforementioned methods (Rombach et al., 2022; Deng et al., 2022; Zhang et al., 2023; Sohn et al., 2024) are limited due to the following reasons: 1) Lacking the style reference imagery for various domain; 2) Being prone to altering the original semantic content of the images, such as shifts in the positions of small objects, deformations of large objects, and distortions of edge contours, in which preserving the geo-spatial information of semantic images is vital for environmental monitoring and disaster response; 3) Time-consuming and computation capacity-consuming for aerial image at 512×512 resolution) or higher.

1.3 Essence and Contributions of this Work

To address the challenges of domain adaptation in current aerial image segmentation, we proposed the **L**atent **A**erial **S**tyle **T**ransfer model (**LAST**). This model transfers domain information from synthetic data, generated using a game engine, to enhance real aerial images. Specifically, we first utilize a VAE encoder to simultaneously compress both the style reference image and the semantic content image into latent space. The interaction between the style and content is then processed through transformer blocks in this latent space. Finally, the transformed output is decoded back into image scale using the VAE decoder.

To complement existing datasets and address their limitations, we developed the **A**erial **W**eather **S**ynthetic **D**ataset (**AWSD**), which introduces con-

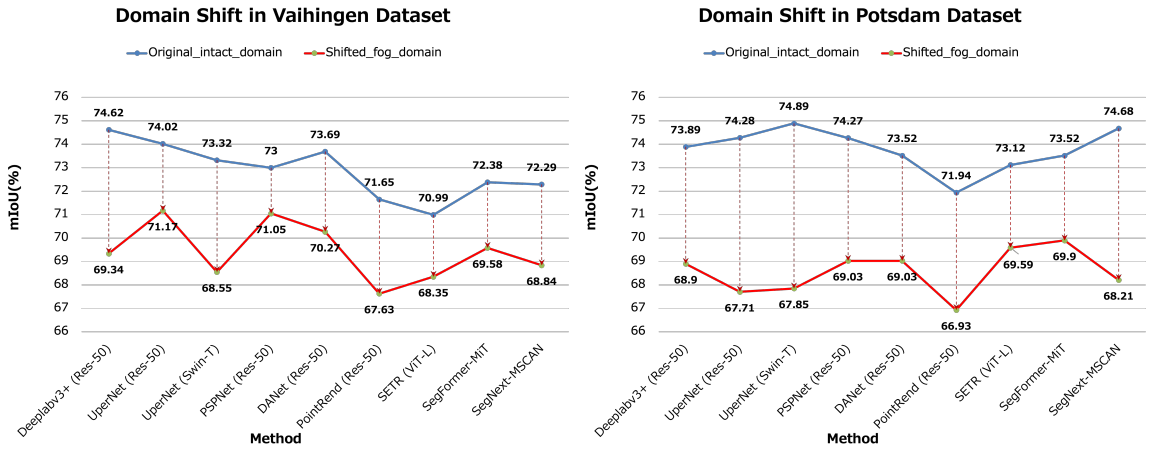

Figure 2: **Domain Shift** in ISPRS (Gerke, 2012; Rottensteiner et al., 2014) Vaihingen (*left*) and Potsdam (*right*) dataset. As the first step, we pre-trained 9 prevalent Segmentors: Deeplabv3+(Chen et al., 2018), UperNet(Xiao et al., 2018), PSPNet(Zhao et al., 2017), DANet(Fu et al., 2019), SETR(Zheng et al., 2021), SegFormer(Xie et al., 2021), SegNext(Guo et al., 2022), PointRend(Kirillov et al., 2020) with varied backbones: ResNet(He et al., 2016), ViT(Dosovitskiy et al., 2020), SwinTransformer(Liu et al., 2021) in original training set under the same setting. Afterwards, we tested them on intact validation set (blue-curve in figure) and our generated fog validation set (green, red-curve in figure) respectively. The preliminaries demonstrates the model performance deterioration caused by domain shift from original to fog condition. Precisely **-3.46% mIoU** in Vaihingen and **-5.22% mIoU** in Potsdam respectively. Zoom in for the best view.

trolled variations in weather and lighting. This dataset provides an ideal benchmark for evaluating the robustness of segmentation models in diverse environmental conditions. Leveraging this dataset, we generated realistic foggy domain data, which supplements existing aerial image segmentation datasets like ISPRS Vaihingen and Potsdam (Gerke, 2012; Rottensteiner et al., 2014).

We focused specifically on foggy weather, a typical domain shift scenario where dense fog reduces illumination and obscures scene elements. This allowed us to demonstrate the effects of domain shift and present domain adaptation results step by step. In summary, our work contributes the following:

1) We developed **AWSD** using a game engine (Unreal, 2024), offering a variety of domain conditions (e.g., fog, snow, night) to tackle the scarcity of domain-specific data in aerial image segmentation.

2) We introduced **LAST**, a style transfer model that operates in latent space, enabling the transformation of synthetic styles into existing ISPRS aerial datasets (Gerke, 2012; Rottensteiner et al., 2014).

3) We benchmarked state-of-the-art segmentation models on multi-domain datasets generated via AWSD and LAST. Extensive experiments reveal the performance degradation caused by domain shifts, and we successfully adapted model performance in the shifted domain while maintaining its effectiveness in the source domain.

2 RELATED WORK

2.1 Semantic Segmentation

Following the pioneer approach, i.e., Fully Convolutional Network (FCN) (Long et al., 2015), encoder-decoder structure has been a prevalent paradigm for semantic segmentation task. In the early stage, these methods (Ronneberger et al., 2015; Badrinarayanan et al., 2017; Zhao et al., 2017; Lin et al., 2017a) combined the low level feature and its up-sampling high level to obtain the precise objects boundaries meanwhile capture the global information. Consequently, deeplab-series methods(Chen et al., 2017a; Chen et al., 2017b) developed the dilated convolutions to enlarge the receptive field of convolutional layers and further employed spatial pyramid pooling modules to obtain multi-level aggregated feature.

In addition to CNN-based semantic segmentation methods, vision transformer-based approaches(Dosovitskiy et al., 2020; Liu et al., 2021; Fu et al., 2019; Guo et al., 2022) have also become popular due to their exceptional ability to capture long-range contextual information among tokens or embeddings. SETR(Zheng et al., 2021) employs ViT as its backbone and utilizes a CNN decoder to frame semantic segmentation as a sequence-to-sequence task. Moreover, Segmentor (Strudel et al., 2021) introduces a point-wise linear layer following the

ViT backbone to generate patch-level class logits. Additionally, SegFormer(Xie et al., 2021) proposed a novel hierarchically structured Transformer encoder which outputs multiscale features and a MLP decoder to combine both local and global information. Notably, many recent Feature Pyramid Network (FPN)(Lin et al., 2017b)-based affinity learning methods(Xiao et al., 2018; Zheng et al., 2020; Li et al., 2021; Wang et al., 2024a) are proposed to achieve better feature representation and successfully handle the scale-variation problem (Xia et al., 2018; Waqas Zamir et al., 2019) in aerial image segmentation.

2.2 Image Style Transfer

Image Style transfer(Gatys et al., 2016; Johnson et al., 2016; Li and Wand, 2016; Zhu et al., 2017) is practical research field that apply the style of one reference image into the content of another image, it aims to generate a transferred image that contained the content, such as shapes structures, objects of the original content image but adopt the style, such as colors, textures, patterns of the reference style image. The pioneer methods (Gatys et al., 2016) demonstrates that CNNs' hierarchical layers can extract content and style information, proposing an optimization-based method for iterative stylization. However, the network is usually limited to a fixed set of styles and cannot adapt to arbitrary new styles. To fix the deficiency of the previous, AdaIN-style(Huang and Belongie, 2017) presents a novel adaptive instance normalization (AdaIN) layer that aligns the mean and variance of the content features with those of the style features. (Chen et al., 2021) deploy a internal-external learning scheme with two types of contrastive loss, which can make generated image more reasonable and harmonious. StyTr^2 (Deng et al., 2022) is the first baseline for style transfer using a visual transformer, that is capable of domain-specific long-range information. Despite that, the inference computation speed is inferior to the CNN-based approaches.

2.3 Domain Shift

Domain shift (Ben-David et al., 2010) is a well-known challenge that results in unforeseen performance degradation under conditions different from those in the training phase. To address this issue, domain generalization (Khosla et al., 2012; Muandet et al., 2013; Tobin et al., 2017; Volpi et al., 2018) based algorithm has been developed to generalize learning model across weather conditions and city environments unexplored during training (target samples are not available during training). In addition to that, a sub-field of transfer learning, i.e., domain adaptation-based methods are also proposed to adapt a model trained on data from the source domain to perform effectively on data from target domain (Tzeng et al., 2017; Wang and Deng, 2018; Farahani et al., 2021). Generally, domain adaptation algorithm aims to learn a model from a source labeled data that can be extended to a target domain by minimizing the difference between domain distributions.

The exploration of domain shift solutions largely depends on the availability of target domain data, which is often rare and difficult to acquire, especial for the diverse weather conditions. Moreover, annotating data for new domains is a laborious and time-consuming task. Therefore, unlike the aforementioned methods, we utilized Unreal Engine(Unreal, 2024) to build a synthetic dataset that encompasses a wide variety of weather conditions (details provided in Section 3.1). On top of this, we applied style transfer to augment the already fine-annotated ISPRS Vaihingen and Potsdam datasets(Rottensteiner et al., 2014). As a result, by performing simple joint training on both the source and shifted domains, we can effectively address domain shift and the accompanying degradation in model performance.

3 APPROACHES

To achieve style transfer for aerial images, accounting for variations in weather conditions and illumination while reducing the computational cost of processing, we propose the **LAST** model. This model operates in two spaces: image space and latent space, as depicted in Figure 3. Specifically, inspired by the Latent Diffusion Models (LDMs(Rombach et al., 2022)), we first compress the input aerial images into the latent space using a pre-trained VAE (Section 3.1). The style transformation is then performed in this latent space (Section 3.2). Additionally, the perceptual loss(Johnson et al., 2016), computed via a pre-trained VGG-19(Simonyan and Zisserman, 2014), is applied to optimize the model (Section 3.3).

3.1 VAE for Image Compression

We first deploy the same setup as Latent Diffusion(Rombach et al., 2022) to compress image into the latent space via the variational autoencoder (VAE(Kingma, 2013; Vahdat and Kautz, 2020)) pre-trained under the Kullback-Leibler (KL) Divergence penalty.

Given an image $x \in \mathbb{R}^{H\times W\times 3}$ in image space, the

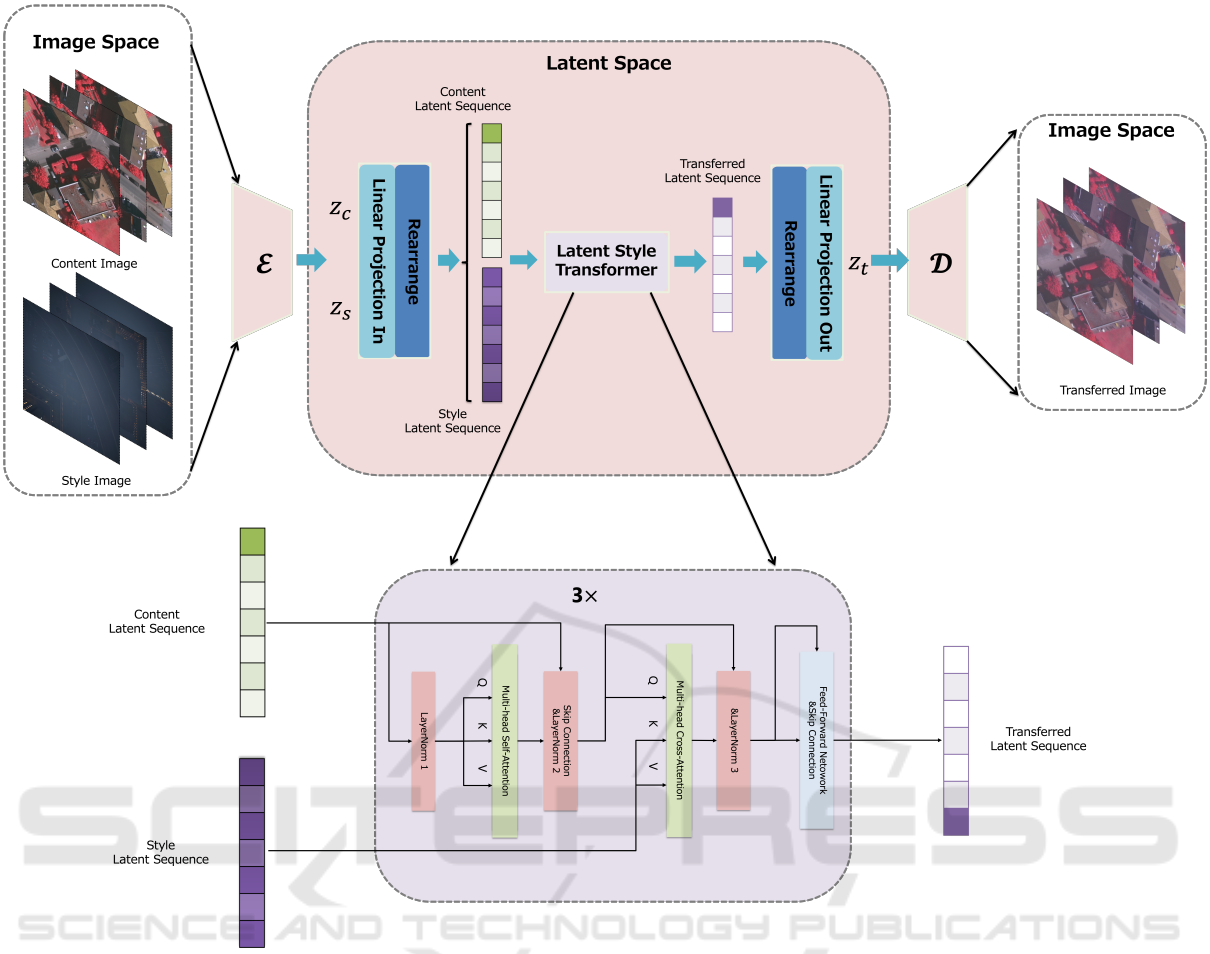

Figure 3: **Pipeline of the LAST**. Given the image pairs Content images, Style images in image space, they are first compressed into latent space via the VAE encoder, then flatten into the latent sequences. Afterwards the style transfer interaction are implemented in the latent style transformer. Finally, the latent outputs are recovered into image space and generate the transferred images.

encoder $\mathcal{E}$ encodes x into a latent representation $z \in \mathbb{R}^{h\times w\times C}$, where the $h = H/f$, $w = W/f$ and the down-sampling factors $f = 4$. Afterwards, the decoder $\mathcal{D}$ decodes the latent vector z to obtain the reconstruct image $\tilde{x} = \mathcal{D}(z)$. Specifically, it primarily contains the following processes in the **LAST**:

- An endoder $\mathcal{E}$ to encodes the input content and style image pair $[x_c, x_s] \in \mathbb{R}^{H\times W\times 3}$ into two Gaussian distributions:

$$\mathcal{N}(\mu_c, \sigma_c^2) = \mathcal{E}(x_c) \tag{1}$$

$$\mathcal{N}(\mu_s, \sigma_s^2) = \mathcal{E}(x_s) \tag{2}$$

- Adopting reparameterization trick(Kingma, 2013; Figurnov et al., 2018) to sample the latent vector z_c and z_s respectively from the encoded Gaussian distributions. In particular,

$$z_c = \mu_c + \sigma_c \odot \varepsilon \tag{3}$$

$$z_s = \mu_s + \sigma_s \odot \varepsilon \tag{4}$$

where "$\odot$" denotes element-wise multiplication, $\varepsilon \sim \mathcal{N}(0,1)$ and $[z_c, z_s] \in \mathbb{R}^{h\times w\times C}$

- Within the latent space, the latent vector $[z_c, z_s]$ are projected into sequence and interacted with each other in Latent Style Transformer (**LSTrans**), which outputs the transferred sequence and projects it out to latent vector as follows:

$$z_t = LSTrans(z_c, z_s) \tag{5}$$

where, $z_t \in \mathbb{R}^{h\times w\times C}$. Finally the VAE decoder $\mathcal{D}$ decodes out the style transferred image $x_t = \mathcal{D}(z_t)$, where $x_t \in \mathbb{R}^{H\times W\times 3}$.

3.2 Latent Style Transformer

In this section, we introduce the proposed Latent Style Transformer (**LSTrans**). The detailed pipeline is il-

lustrated in Figure 3. The latent vectors, denoted as $z \in \mathbb{R}^{h\times w\times C}$, are first flattened and embedded into latent sequences, represented in $s \in \mathbb{R}^{hw\times C}$. To transfer domain-specific information from the input style image to the content image while preserving original semantic details—such as objects, boundaries, and spatial relationships—we stack three sequential transformer blocks in the latent space to process the compressed latent vectors. Each block consists of the following components:

- The first Multi-head Self-Attention (**MSA**) to grasp the contextual information for content images.
- The second Multi-head Cross-Attention (**MCA**) to facilitate interaction between flattened latent vectors.
- The last Feed-Forward Network (**FFN**) to enhance the model's capacity for non-linear transfomration and feature combination.

As a result, **LSTrans** outputs the transferred latent sequence, after rearrange and reversed embedding, we obtain the transferred latent vector $z_t \in \mathbb{R}^{h\times w\times C}$, which is decoded into $x_t \in \mathbb{R}^{H\times W\times 3}$ in image space.

3.3 Perceptual Loss for Model Optimization

To obtain the transferred image x_t that keep the content of the x_c while containing the style of the x_s, following the previous style transfer approaches(Johnson et al., 2016; Huang and Belongie, 2017; Chen et al., 2021; Deng et al., 2022), we import Perceptual loss (VGG loss) as the penalty at each training step. The total loss is defined as:

$$L = w_1 L_c + w_2 L_s \tag{6}$$

in which Lc and Ls respectively compute the content loss between x_t and x_c, style loss between x_t and x_s, w_1 and w_2 are weighted factors and set to 1 and 0.8. Given the pre-trained VGG-19 and input image $x \in \mathbb{R}^{H\times W\times 3}$, the first four convolutional layers output are the low-level features f_l that denote the image style and domain information, while the last two convolutional layers output are the high-level features f_h that denotes the image's smeantic content. Thus the content style loss and content loss is compute as follows:

$$L_c = \|f_{ht} - f_{hc}\|^2 \tag{7}$$

$$L_s = \|f_{lt} - f_{ls}\|^2 \tag{8}$$

here, f_{ht}, f_{hc}, f_{lt}, and f_{ls} respectively represent, the high-level features of x_t, the high-level features of x_c, the low-level features of x_t, and the low-level features of x_s.

4 EXPERIMENTS

4.1 Datasets

Existing aerial image segmentation datasets, such as ISPRS Potsdam and Vaihingen(Gerke, 2012; Rottensteiner et al., 2014), serve as widely-used benchmarks, offering high-resolution, annotated images of urban environments. While these datasets are invaluable for training and evaluating segmentation models, they have significant limitations in real-world applications. A key issue is their lack of diversity in environmental conditions. Both datasets primarily feature images captured under ideal circumstances, such as clear skies and uniform lighting, which do not accurately reflect the variability present in real-world aerial imagery(Wang et al., 2024a). Consequently, models trained on these datasets often struggle with domain shifts—environmental changes like weather or lighting variations that can drastically reduce segmentation accuracy.

In real-world scenarios, such as disaster response or urban planning, aerial images are frequently taken under challenging conditions, including fog, rain, snow, or at night. The absence of such environmental diversity in standard datasets limits the robustness and adaptability of segmentation models when deployed in dynamic environments. To address this shortcoming, there is a need for a new dataset that not only mirrors the spatial characteristics of datasets like ISPRS but also includes diverse weather conditions to simulate domain shifts.

4.1.1 ISPRS Dataset

The International Society for Photogrammetry and Remote Sensing (ISPRS) Vaihingen and Potsdam datasets (Gerke, 2012; Rottensteiner et al., 2014) are two widely used benchmarks from the ISPRS 2D Semantic Labeling Contest. The Vaihingen dataset consists of high-resolution aerial images of Vaihingen, Germany, captured as true orthophotos with a ground sampling distance (GSD) of 9 cm. It includes 33 image tiles, 16 of which are annotated with six semantic categories: *impervious surfaces, buildings, low vegetation, trees, cars, and clutter (background)*. The Potsdam dataset provides aerial images of Potsdam, Germany, captured with a finer GSD of 5 cm. It contains 38 tiles, each depicting diverse urban and suburban landscapes, and is similarly annotated with six semantic classes.

The original ISPRS images are augmented into 512×512 small images through cropping operations. Afterwards, the augmented images are configured for benchmarking, where 3,456 images for training and

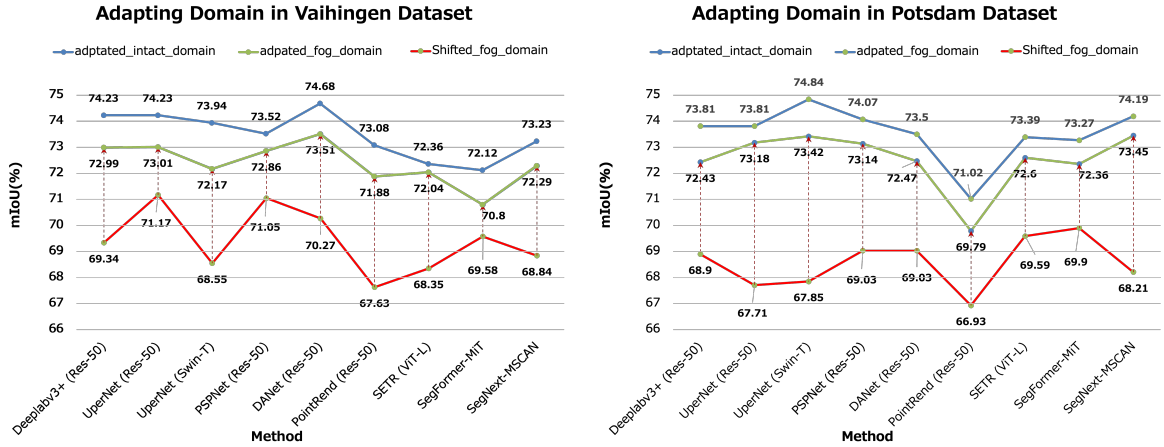

Figure 4: **Results of Adapting to Foggy Domains** in the ISPRS (Gerke, 2012; Rottensteiner et al., 2014) Vaihingen (*left*) and Potsdam (*right*) datasets. The green and red curve denotes results on foggy validation set of segmentors train with and w/o shifted foggy domain data. Training with foggy domain data resulted in segmentors achieving an increase of **+2.97% mIoU** in Vaihingen and **+3.97% mIoU** in Potsdam, compared to training solely in the original domain. Moreover, the blue curve shows the models trained with shifted domain data still keep their capacity on original domain. Zoom in for optimal viewing.

2,016 images for validation in Potsdam and 344 images for training and 398 images for validation in Vaihingen. To demonstrate the effect of domain shift, we train the 9 segmentation models (Deeplabv3+(Chen et al., 2018), UperNet(Xiao et al., 2018)-Res50(He et al., 2016), UperNet-SwinT(Liu et al., 2021), PSPNet(Zhao et al., 2017), DANet(Fu et al., 2019), SETR(Zheng et al., 2021), SegFormer(Xie et al., 2021), SegNext(Guo et al., 2022), PointRend(Kirillov et al., 2020)) under the original ISPRS training set and test them under the intact ISPRS validation set and their style-transferred foggy version, the results is illustrated in Figure 2.

4.1.2 Synthetic Dataset

AWSD is a synthetic dataset created using Unreal Engine 5 (Unreal, 2024), designed to replicate realistic urban environments modeled after the Potsdam and Vaihingen datasets. The dataset captures images from a 200-meter aerial perspective, maintaining consistency with original benchmarks in terms of viewpoint and object layout.

In contrast to the static, clear-sky images in ISPRS datasets(Gerke, 2012; Rottensteiner et al., 2014), AWSD includes diverse weather conditions such as snow, fog, and nighttime scenes. These conditions were purposefully introduced to assess the adaptability of segmentation models to domain shifts. AWSD retains the same pixel-level semantic annotations across six urban categories as ISPRS, ensuring precise training and testing for both small and large objects in complex environments.

By introducing varied weather scenarios, AWSD addresses the challenge of domain shifts, enabling models to generalize more effectively across different conditions. Its synthetic nature allows for the consistent simulation of environmental variations that are hard to capture in real-world datasets, making it a valuable resource for enhancing aerial segmentation algorithms' robustness in real-world applications.

4.2 Comparison Study

4.2.1 Preliminary Setting

To demonstrate the adverse effects of domain shift and simultaneously generate a foggy domain dataset, we first trained the **LAST** model using the ISPRS dataset alongside foggy images from the **AWSD**. We combined the training sets of Vaihingen and Potsdam as the content image set, while 462 synthetic foggy images from UE5 (Unreal, 2024) served as the style image set. Using Adam as the optimizer, we trained the latent style transformer for 32,000 iterations on two Nvidia RTX 3090 GPUs. During this process, the parameters in both the VAE and the perceptual VGG-19 models were kept frozen.

4.2.2 Experimental Results

The well-trained **LAST** model is used to generate foggy versions of the ISPRS training and validation sets. This eliminates the need for additional domain adaptation algorithms or extra annotation efforts for the foggy domain data. Furthermore, we combine the original ISPRS training set with its foggy version and re-evaluate the performance of nine different segmen-

Table 1: **Comparison experiment on the Vaihingen dataset**. We evaluate the performance of segmentors on the original domain validation set, comparing results from training without (w/o.) and with (w.) shifted domain data.

Method	mIoU(%) w/o. shifted data ↑	mIoU(%) w. shifted data ↑
Deeplabv3+(Chen et al., 2018) (Res-50)	74.62	74.23
UperNet(Xiao et al., 2018) (Res-50)	74.02	74.23
UperNet (Swin-T)	73.32	73.94
PSPNet(Zhao et al., 2017) (Res-50)	73.00	73.52
DANet(Fu et al., 2019) (Res-50)	73.69	74.68
PointRend(Kirillov et al., 2020) (Res-50)	71.65	73.08
SETR(Zheng et al., 2021) (ViT-L)	70.99	73.90
SegFormer(Xie et al., 2021) (MiT)	72.38	72.12
SegNext(Guo et al., 2022) (MSCAN)	72.29	73.23
Average	72.88	**73.49**

Table 2: **Comparison experiment on the Potsdam dataset**. We evaluate the performance of segmentors on the original domain validation set, comparing results from training without (w/o.) and with (w.) shifted domain data.

Method	mIoU(%) w/o. shifted data ↑	mIoU(%) w. shifted data ↑
Deeplabv3+(Chen et al., 2018) (Res-50)	73.89	73.81
UperNet(Xiao et al., 2018) (Res-50)	74.28	73.81
UperNet (Swin-T)	74.89	74.84
PSPNet(Zhao et al., 2017) (Res-50)	74.27	74.07
DANet(Fu et al., 2019) (Res-50)	73.52	73.50
PointRend(Kirillov et al., 2020) (Res-50)	71.94	71.02
SETR(Zheng et al., 2021) (ViT-L)	73.12	73.39
SegFormer(Xie et al., 2021) (MiT)	73.52	73.27
SegNext(Guo et al., 2022) (MSCAN)	74.68	74.19
Average	**73.79**	73.54

tation models. The results are presented in Figure 4 and Tables 1 and 2.

We evaluate the performance of the nine segmentation models on the ISPRS validation set and its foggy domain counterpart. Without requiring any additional annotations for the shifted domain, the robustness of all models to the foggy domain improved by **+2.97% mIoU** on Vaihingen and **+3.97% mIoU** on Potsdam. Moreover, as shown in Tables 1 and 2, their performance in the original domain (clear weather with adequate lighting conditions) is preserved.

5 CONCLUSIONS

In this work, we employ synthetic image style transfer to address domain shifts in aerial imagery. First, we developed the Aerial Weather Synthetic Dataset (**ASWD**), which introduces rare domain conditions from an aerial perspective. Additionally, we proposed a Latent Aerial Style Transfer model (**LAST**) to transform original aerial data into a foggy version—a typical domain shift that alters weather and lighting conditions. This approach eliminates the need for additional annotations in the foggy domain. Finally, we benchmarked current state-of-the-art (*SoTA*) segmentation methods on the foggy ISPRS dataset, highlighting the impact of domain shift and successfully adapting model performance to the new domain while maintaining its effectiveness in the original domain.

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79:151–175.

Brock, A. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Brooks, T., Holynski, A., and Efros, A. A. (2023). Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402.

Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chen, H., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., Lu, D., et al. (2021). Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34:26561–26573.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017a). 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.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017b). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

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

Chung, J., Hyun, S., and Heo, J.-P. (2024). Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8795–8805.

Dai, D. and Van Gool, L. (2018). Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE.

Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., and Xu, C. (2022). Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. (2021). A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894.

Figurnov, M., Mohamed, S., and Mnih, A. (2018). Implicit reparameterization gradients. *Advances in neural information processing systems*, 31.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., and Lu, H. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3146–3154.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423.

Gerke, M. (2012). Use of the isprs vaihingen and potsdam datasets for urban classification analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*. ISPRS WG III/4.

Gong, M., Xie, S., Wei, W., Grundmann, M., Batmanghelich, K., Hou, T., et al. (2024). Semi-implicit denoising diffusion models (siddms). *Advances in Neural Information Processing Systems*, 36.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z., Cheng, M.-M., and Hu, S.-M. (2022). Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 770–778.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer.

Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., and Laine, S. (2024). Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24174–24184.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Khanna, S., Liu, P., Zhou, L., Meng, C., Rombach, R., Burke, M., Lobell, D. B., and Ermon, S. (2023). Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., and Torralba, A. (2012). Undoing the damage of dataset bias. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 158–171. Springer.

Kingma, D. P. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kirillov, A., Wu, Y., He, K., and Girshick, R. (2020). Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer*

vision and pattern recognition (CVPR), pages 9799–9808.

Li, C. and Wand, M. (2016). Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 702–716. Springer.

Li, T., Chang, H., Mishra, S., Zhang, H., Katabi, D., and Krishnan, D. (2023). Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2142–2152.

Li, X., He, H., Li, X., Li, D., Cheng, G., Shi, J., Weng, L., Tong, Y., and Lin, Z. (2021). Pointflow: Flowing semantics through points for aerial image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4217–4226.

Liang, Y., Li, X., Tsai, B., Chen, Q., and Jafari, N. (2023). V-floodnet: A video segmentation system for urban flood detection and quantification. *Environmental Modelling & Software*, 160:105586.

Lin, G., Milan, A., Shen, C., and Reid, I. (2017a). Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017b). Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2117–2125.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3431–3440.

Luo, Z., Gustafsson, F. K., Zhao, Z., Sjölund, J., and Schön, T. B. (2023). Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1680–1691.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M., and Brendel, W. (2019). Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*.

Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR.

Peebles, W. and Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

Pi, Y., Nath, N. D., and Behzadan, A. H. (2020). Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Advanced Engineering Informatics*, 43:101009.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J. D., Breitkopf, U., and Jung, J. (2014). Results of the isprs benchmark on urban object detection and 3d building reconstruction. *ISPRS journal of photogrammetry and remote sensing*, 93:256–271.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France. PMLR.

Sohn, K., Jiang, L., Barber, J., Lee, K., Ruiz, N., Krishnan, D., Chang, H., Li, Y., Essa, I., Rubinstein, M., et al. (2024). Styledrop: Text-to-image synthesis of any style. *Advances in Neural Information Processing Systems*, 36.

Song, J., Meng, C., and Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272.

Sun, T., Segu, M., Postels, J., Wang, Y., Van Gool, L., Schiele, B., Tombari, F., and Yu, F. (2022). Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE.

Toker, A., Eisenberger, M., Cremers, D., and Leal-Taixé, L. (2024). 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.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176.

Unreal, E. (2024). https://www.unrealengine.com/en-us.

Vahdat, A. and Kautz, J. (2020). Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. (2018). Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31.

Wang, M. and Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Wang, Y., Wang, Z., Nakano, Y., Hasegawa, K., Ishii, H., and Ohya, J. (2024a). Mac: Multi-scales attention cascade for aerial image segmentation. In *13th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2024*, pages 37–47. Science and Technology Publications, Lda.

Wang, Y., Wang, Z., Nakano, Y., Nishimatsu, K., Hasegawa, K., and Ohya, J. (2022). Context enhanced traffic segmentation: traffic jam and road surface segmentation from aerial image. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE.

Wang, Z., Jiang, Y., Zheng, H., Wang, P., He, P., Wang, Z., Chen, W., Zhou, M., et al. (2024b). Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in neural information processing systems*, 36.

Wang, Z., Zhao, L., and Xing, W. (2023). Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7677–7689.

Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.-S., and Bai, X. (2019). isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 28–37.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. (2018). Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. (2018). Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPs)*, 34:12077–12090.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915.

Zhang, L., Rao, A., and Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2881–2890.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890.

Zheng, Z., Zhong, Y., Wang, J., and Ma, A. (2020). Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4096–4105.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.