2024 21st International Conference on Ubiquitous Robots (UR)
June 24 - 27, 2024, New York University, USA

QuadFormer: Real-Time Unsupervised Power Line Segmentation with Transformer-Based Domain Adaptation

Pratyaksh Prabhav Rao[1*], Feng Qiao[2*], Weide Zhang[3], Yiliang Xu[4], Yong Deng[4], Guangbin Wu, Qiang Zhang, and Giuseppe Loianno[1]

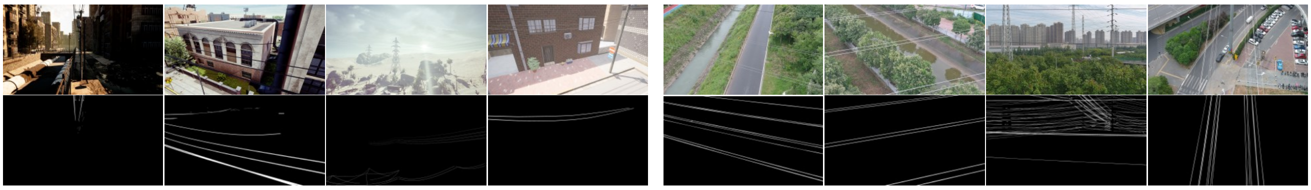

Fig. 1: AutelPL Dataset: Examples from AutelPL Synthetic (left), AutelPL Real (right), and ground truth (bottom).

2024 21st International Conference on Ubiquitous Robots (UR) | 979-8-3503-6107-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/UR61395.2024.10597474

Abstract— **Accurately identifying Power Lines (PLs) is crucial for ensuring the safety of aerial vehicles. Despite the potential of recent deep learning approaches, obtaining high-quality ground truth annotations remains a challenging and labor-intensive task. Unsupervised Domain Adaptation (UDA) emerges as a promising solution, leveraging knowledge from labeled synthetic data to improve performance on unlabeled real images. However, existing UDA methods often suffer of huge computation costs, limiting their deployment on real-time embedded systems commonly utilized on aerial vehicles. To mitigate this problem, this paper introduces QuadFormer, a real-time framework designed for unsupervised semantic segmentation within the UDA paradigm. QuadFormer integrates a lightweight transformer-based segmentation model with a cross-attention mechanism to narrow the gap between a labelled synthetic domain and unlabelled real domain. Furthermore, we design a novel pseudo label scheme to enhance the segmentation accuracy of the unlabelled real data. To facilitate the evaluation of our framework and promote reserach in PL segemntation, we present two new datasets: AutelPL Synthetic and AutelPL Real. Experimental results demonstrate that QuadFormer achieves state-of-the-art performance on both AutelPL Synthetic → TTPLA and AutelPL Synthetic → AutelPL Real tasks. We will publicly release the dataset to the research community.**

Supplementary Material

Video: https://youtu.be/7h-lqGbQCSg

I. Introduction

Unmanned Aerial Vehicles (UAVs) have gained widespread usage in diverse fields such as photography, commercial, agriculture, and exploration. One of the major challenges during UAV flight is their inability to detect widespread PLs. Collision with PLs would not only damage the UAV but can adversely affect power grids. Therefore, the development of accurate PL detection methods assumes paramount importance. However, the segmentation of PLs presents several challenges. Their geometric structure is thin, occupying only a tiny portion of the image, making detection intricate. Additionally, the presence of cluttered backgrounds with similar-looking edges, low contrast scenarios, or barely visible thickness further complicates their identification. Recently, numerous deep learning methods have emerged for semantic segmentation and shown promising results [1]–[3]. Particularly, transformer-based models [4]–[6] have achieved remarkable generalization performance on segmentation tasks. However, these methods demand heavy computational resources, posing practical deployment challenges, especially for real-time applications. Real-time segmentation of PLs is crucial to provide drones with sufficient time to avoid potential collisions effectively. Furthermore, these approaches rely on extensive annotated datasets, which are not easily available. The process of collecting them can also be cumbersome and expensive. Particularly, there is a scarcity of high-quality open-source datasets with labeled PL segmentation data. Hence, finding efficient and resource-friendly solutions for PL detection becomes essential to ensure the safe operation of UAVs.

UDA presents a promising approach to address the challenges outlined above. This technique aims to bridge the domain gap between a labeled source domain and an unlabeled target domain. Although potentially promising, existing UDA methods often face limitations in real-time inference capabilities. These techniques often utilize large models that are unable to run on real-time systems commonly utilized by UAVs. Our research seeks to overcome this critical limitation by proposing and evaluating a real-time UDA framework designed to enhance the precision of PL detection on embedded systems. The contributions can be summarized as follows. First, we adapt a lightweight SegFormer model [5] for robust real-time PL detection without the need for annotated data, addressing the problem within the context of unsupervised domain adaptation. Second, our pro-

[*]These authors contributed equally.

[1]The authors are with the New York University, Tandon School of Engineering, Brooklyn, NY 11201, USA. email: {pr2257, as16054, loiannog}@nyu.edu.

[2]The author is with RWTH Aachen University, Germany

[3]The author is with Coupang Inc., CA, USA

[4]The authors are with Autel Robotics, CA, USA

This work was supported by the NSF CAREER Award 2145277, Qualcomm Research, Nokia, and NYU Wireless. This work has been partially conducted at Autel Robotics during Pratyaksh's internship in summer 2022.

979-8-3503-6107-0/24/$31.00 ©2024 IEEE

TABLE I: Comparison of major powerline segmentation datasets with AutelReal and AutelSynthetic dataset

Dataset	**Images#**	**Resolution**	**Diversity**			**Manual Annotation**	**Synthetic**	**Real**
			Zoom Level	Illumination	Angle			
PLID [7]	2000	128 × 128	✓	✗	✓	✗	✗	✓
PLDU [8]	573	540 × 360	✓	✗	✓	✓	✗	✓
PLDM [8]	287	540 × 360	✓	✗	✗	✓	✗	✓
TTPLA [9]	1100	3840 × 2160	✓	✓	✓	✓	✗	✓
AutelReal (ours)	4000	3840 × 2160	✓	✓	✓	✓	✗	✓
WDD [10]	67000	640 × 480	✓	✓	✓	–	✓	✗
AutelSynthetic (ours)	7000	3840 × 2160	✓	✓	✓	–	✓	✗

posed framework integrates cross-attention and self-attention mechanisms, facilitating the learning of robust feature representations. Third, we enhance detection precision by incorporating self-training [11] and pseudo-label correction [12] mechanisms. This novel pseudo-label scheme estimates the reliability of the pseudo ground truth labels for the target domain and denoises them by leveraging class representations learned using the two attention mechanisms. Finally, to advance research in PL segmentation, we introduce two publicly available datasets: AutelPL Synthetic, featuring 4K resolution images with annotations from AirSim [13], and AutelPL Real, comprising 4K resolution images obtained from flight videos with precise ground truth annotations. Compared to many existing datasets, which often suffer from limitations such as a limited number of images, sparse annotations, and low resolution (see Table I), our datasets offer significant advantages. They address the scarcity of high-quality data for PL segmentation and also ensure that our models are well-equipped to handle various real-world challenges and operating conditions effectively.

II. Related Works

A. Powerline Segmentation

The landscape of PL detection methods has evolved over time, transitioning from early utilization of traditional computer vision techniques to the recent advancements in deep learning-based approaches. Earlier works, exemplified by [14], employ sub-pixel edge detection combined with thresholded Hough transform, and [15] applies a Canny detector with subsequent post-processing refinement. These methods rely on handcrafted features and predefined hyperparameters, limiting their adaptability to diverse aerial conditions. Conversely, contemporary methods leverage deep learning for PL segmentation. [16] partitions images into sub-regions, classifying patches with PLs. [17] and [18] investigate the effectiveness of Convents for PL detection. [19] introduced a two-phase weakly supervised detection method for PL extraction. Their approach involved a Convolutional Neural Network (CNN) for approximate positioning and a subsequent refinement algorithm to connect broken lines. While general CNN-based semantic segmentation models [1] exhibit promise in various segmentation tasks, they fall short for PL segmentation due to information loss during pooling and downsampling. In contrast, transformer-based models [4]–[6], are recognized for their proficiency in leveraging long-term spatial contextual dependencies, resulting in state-of-the-art performance. However, they rely heavily on substantial annotated datasets and are computationally expensive. In this work, we adapt a lightweight transformer-based model, SegFormer-b0 [5], for PL segmentation in the context of unsupevised domain adaptation, circumventing the need for manual annotations.

B. Poweline Segmentation Datasets

Generating segmentation datasets with pixel-level annotations is a laborious and costly endeavor. This challenge is further compounded for PL detection due to the inherent difficulty in capturing images in close proximity to power grids while UAVs are in flight [18]. Despite these difficulties, a handful of publicly available datasets for PL detection do exist. For instance, the PLID dataset [7] comprises 2000 low-resolution images featuring PLs. Similarly, [8] contribute two datasets, PLDU and PLDM, covering urban and mountain scenarios. However, these datasets still have limitations in terms of diversity, quantity, and image resolution of annotated PLs. Addressing these gaps, the TTPLA dataset [9] introduces 1100 4K resolution images with meticulous annotations. Nevertheless, this dataset is subject to bias stemming from city-specific image acquisition. A promising alternative is to use synthetic images. The wire detection dataset (WDD) [10] overlays synthetic wires onto 67K aerial images. Yet, a notable drawback lies in the non-photorealistic quality of the synthetic wires. In this work, we present a novel synthetic dataset, AutelPL Synthetic, which simulates high-quality urban scenes with diverse PL scenarios and automatically generated annotations. Furthermore, to advance PL segmentation research, we introduce AutelPL Real, a real-world PL segmentation dataset comprising 4000 4K resolution images showcasing PL instances. Both datasets offer a broad range of images with significant variability, encompassing changes in illumination, textures, and camera perspectives.

C. Unsupervised Domain Adaptation

Recent UDA developments are primarily divided into two categories: adversarial training, aligning source and target domain distributions using techniques like [20]–[22], and self-training, which labels target domain data, either precomputed offline and retrained [23] or dynamically during training with adaptive thresholds [24]. Regularization methods, including pseudo-label prototypes [12] and consistency regularization [25], enhance training. DACS [26] combines domain data with labels and pseudo-labels. Transformer-based models, like DAFormer [27] and HRDA [28], have shown promise in UDA, utilizing the Segformer as the base model and specialized training strategies to bridge the domain gap. However, most UDA techniques rely on large models that are impractical for real-time embedded systems. Our work focuses on leveraging cross-domain contextual insights, inspired by BCAT [29], with a lightweight Segformer, achieving real-time and accurate segmentation in the target domain. Differently, we align relative feature distances based on an augmented representation that integrates in-domain and cross-domain contextual information. Moreover, the BCAT method was designed specifically for classification tasks. In this work, we demonstrate the applicability of our approach on a class-imbalanced segmentation task.

III. The Autel Dataset

We introduce two PL datasets: the AutelPL Synthetic and AutelPL Real datasets, each equipped with detailed segmentation annotations. These datasets have been carefully curated to serve as valuable resources for advancing PL segmentation tasks. The AutelPL Synthetic dataset comprises a collection of photorealistic frames generated from four distinct urban scenarios, created using the Unreal Engine 5. Each frame is enriched with pixel-level semantic annotations obtained through the AirSim [13] plugin. To simulate real-world conditions and enhance the complexity of detection tasks, scenes are populated with elements commonly encountered during UAV flights, such as trees, lamp poles, buildings, etc. The virtual environment's flexibility allows for seamless placement and annotation of these elements, facilitating the creation of diverse urban landscapes. Moreover, the ability to manipulate attributes such as textures, colors, and shapes adds visual variety to the dataset. The AutelPL Synthetic dataset includes 7000 frames, each accompanied by ground truth annotations of resolution 3840×2160. These frames are captured from a virtual array of cameras, moving randomly throughout the scene while adhering to a height range of 10 to 15 meters from the ground. Within each camera pose, multiple frames are captured, introducing variations in dynamic objects, scene illumination, and aerial background textures. A minimum separation of 5 m between camera positions ensures enhanced visual diversity.

The AutelPL Real dataset is obtained using a UAV which is flown over different cities, ensuring a mix of scenes. We choose locations within the city randomly to avoid any bias or manipulation of backgrounds. The UAV is equipped with a 4K camera with lossless zoom capabilities. This zooming feature is used while collecting video data to make sure we get detailed images of PLs without needing to manually crop them. All the aerial videos are in 3840×2160 resolution at 30 Hz. We carefully check the dataset and remove any unclear images to maintain its quality. The AutelPL Real dataset consists of 4000 images, each matched with accurate annotations, all at the resolution of 3840×2160. PLs have a distinct appearance, being long and thin, so differences in backgrounds and lighting play a significant role in detecting them. Hence, the dataset covers a range of backgrounds, zoom levels, and various weather and lighting conditions. Understanding the importance of capturing different angles for reliable data collection and ensuring the trained deep learning model is view invariant, we capture videos from various viewpoints. These include front, top, and side views.

IV. Methodology

A. Overview

For a UDA problem, we are given a source dataset $X_s = \{x_s\}_{j=1}^{n_s}$ with ground truth labels $Y_s = \{y_s\}_{j=1}^{n_s}$, and unlabelled target dataset $X_t = \{x_t\}_{j=1}^{n_t}$. The two distributions suffer from a domain shift. The goal of UDA is to train a segmentation model that can provide accurate predictions for X_t. Our training objective can be divided into two stages. First, motivated by the recent success of self-training [30], we generate pseudo labels Y_t^{st} for X_t. Then, during the domain adaptation stage, the segmentation model is retrained with the labelled source dataset, target images, and pseudo labels. Figure 2 illustrates the proposed UDA framework. The QuadFormer consists of a cross-domain transformer encoder (Section IV-B) for generating self-attentive and cross-attentive multi-scale features, and a cross-domain decoder (Section IV-C) for predicting segmentation masks for the source and target domain. Additionally, we implement a pseudo label correction mechanism to online denoise the noisy pseudo ground truth labels.

B. Cross-Domain Transformer Encoder

Segformer [5] has achieved remarkable performance on semantic segmentation. The hierarchically structured transformer encoder outputs multiscale features. Inspired by this, the QuadFormer combines two self-attention and two cross-attention modules to design the four-branch transformer encoder. Given an image pair (one from the source domain and one from the target domain), images are divided into patches of size 4×4. The image patches are transformed into three vectors, namely queries Q, keys K, and values V. The self-attention is formulated as

$$Attn_{self}(Q_s, K_s, V_s) = \text{Softmax}\left(\frac{Q_s K_s^T}{\sqrt{d_{head}}}\right) V_s, \quad (1)$$

$$Attn_{self}(Q_t, K_t, V_t) = \text{Softmax}\left(\frac{Q_t K_t^T}{\sqrt{d_{head}}}\right) V_t, \quad (2)$$

where d_{head} indicates the vector dimension, Q_s, K_s, V_s are queries, keys, and values from the patches of image I_s, and Q_t, K_t, V_t are queries, keys, and values from the patches of image I_t. The cross-attention operation is derived from the self-attention operation. We leverage this module to generate mix-up features and is formulated as

$$Attn_{cross}(Q_s, K_t, V_t) = \text{Softmax}\left(\frac{Q_s K_t^T}{\sqrt{d_{head}}}\right) V_t, \quad (3)$$

$$Attn_{cross}(Q_t, K_s, V_s) = \text{Softmax}\left(\frac{Q_t K_s^T}{\sqrt{d_{head}}}\right) V_s. \quad (4)$$

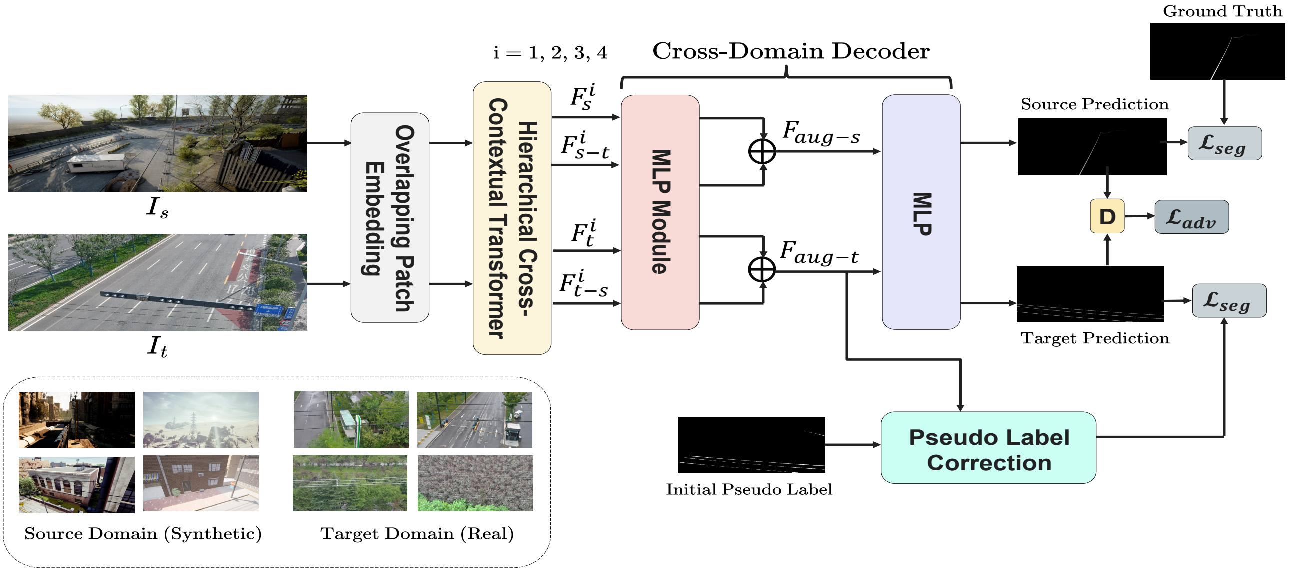

Fig. 2: QuadFormer components include Overlapping Patch Embedding, Hierarchical Cross-Contextual Transformer, Cross-Domain Decoder, and Pseudo Label Correction mechanism.

The proposed hierarchical cross-contextual transformer block consists of four branches: (a) the source-aware branch, (b) the target-aware source branch, (c) the target-aware branch, and (d) the source-aware target branch. The self-attention module extracts hierarchical source-aware features represented as F_s^i and hierarchical target-aware features represented as F_t^i. Moreover, the cross-attention module produces hierarchical target-aware source features denoted as F_{s-t}^i and source-aware target features denoted as F_{t-s}^i. These hierarchical feature maps are expressed as $F^i \in \mathbb{R}^{H/2^{i+1} \times W/2^{i+1} \times C_i}$ with $i = 0, 1, 2, 3$.

C. Cross-Domain Decoder

To achieve consistency in channel dimensions, we process multi-scale features from each branch of the quadruple transformer encoder through an MLP layer, followed by upsampling to ensure resolution harmonization. Subsequently, we enhance the source-aware features and target-aware source features by concatenating the corresponding feature maps to create F_{aug-s}. Similarly, we generate augmented features for the target domain as F_{aug-t}. These augmented features then pass through an additional MLP layer to generate the final segmentation masks namely M$_s$ for the source domain and M$_t$ for the target domain.

D. Pseudo Label Correction

To address self-training challenges, particularly the generation of accurate pseudo-labels for the target dataset, we introduce a dynamic pseudo-label correction mechanism inspired by ProDA [12]. This mechanism estimates pseudo-label reliability by measuring the relative distance between features and representative prototypes, which are centroids of semantic classes. In a departure from traditional prototype calculation methods using target domain features, we propose the generation of cross-attentive prototypes derived from augmented representations, F_{aug-t}. These cross-attentive prototypes amalgamate semantic information from both domains. For each class c, a cross-attentive prototype is computed as a weighted sum of features from the augmented representation for the target domain. These weights are determined by the softmax probability of the corresponding pixel j, as provided by the pseudo-labels. To mitigate the computational burden of prototype calculation, we estimate them as a moving average of semantic cluster centroids within mini-batches. The likelihood of a given pseudo label is estimated by the following weighting scheme

$$W_t^{(j,c)} = \frac{\exp(-||F_{aug-t}^j - \eta^{(c)}||/\tau)}{\sum_{c'} \exp(-||F_{aug-t}^j - \eta^{(c')}||/\tau}, \tag{5}$$

where, $\eta^{(c)}$ is the cross-attentive prototype for a given class c, and τ is the softmax temperature set to $\tau = 1$.

E. Training Objective

The proposed methodology contains a segmentation loss $\mathcal{L}_{seg}$ and an adversarial loss $\mathcal{L}_{adv}$. The segmentation loss of M_s is formulated as

$$\mathcal{L}_{seg}(M_s, Y_s) = -\sum_{i=1}^{H \times W} \sum_{c=1}^{C} Y_s^{i,c} \log M_s^{i,c}, \tag{6}$$

where C is the total number of semantic classes. Similarly, segmentation loss $\mathcal{L}_{seg}(M_t, Y_t^{st})$ is defined for the target segmentation mask. Furthermore, in order to adapt the structured output space [21], we utilize a discriminator to make the source and target masks indistinguishable from each other. To achieve this, we utilize an adversarial loss

$$\mathcal{L}_{adv}(M_s, M_t, D) = \mathbb{E}[\log D(M_s)] + \mathbb{E}[\log 1 - D(M_t)], \tag{7}$$

Therefore, the total loss is

$$\mathcal{L}_{total} = \mathcal{L}_{seg}(M_s, Y_s) + \beta_1 * \mathcal{L}_{seg}((M_t, Y_t^{st})) + \beta_2 * \mathcal{L}_{adv}, \tag{8}$$

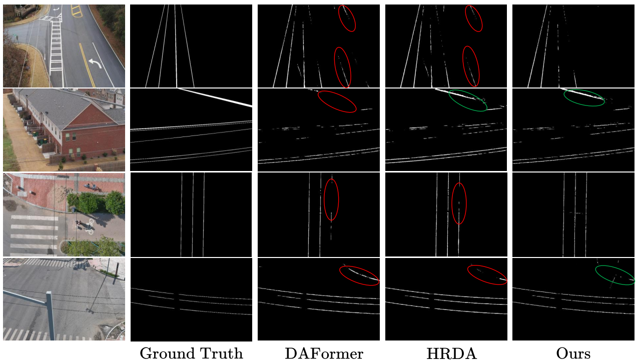

Fig. 3: Qualitative analysis of validation images from TTPLA (first two rows) and AutelPL Real (last two rows), when training models on the AutelPL Synthetic datset. The red circle indicates incorrect predictions and the green circle indicates true positives.

where β_1 and β_2 are set to 0.1, and 1 respectively.

F. Inference for Target Domain

During the inference process on the target distribution after the UDA stage, the inference scheme needs to use the source data. However, there is a storage cost to access the source data. Additionally, it is possible that the source data is not always available to us during inference. Hence, we propose an inference process that is independent of the source data. Since the QuadFormer cannot combine the augmented target feature representation without the source data, we combine the target-aware features with themselves, to predict the segmentation mask M_t during inference.

V. Experiments

A. Implementation Details

Datasets For the target domain, we utilize the AutelPL Real dataset containing 2800 training and 1200 validation images with resolution 3840×2160, and the TTPLA dataset [9] which contains 1004 training images and 217 testing images with resolution 3840×2160. For the source domain, we use the AutelPL Synthetic dataset, which contains 7000 synthetic images of resolution 3840×2160. As a common practice in UDA, we resize the images to 1024×512.

Training Similar to [5], the QuadFormer is trained with AdamW [31], a learning rate of $\eta_{base} = 6 \times 10^{-5}$, a weight decay of 0.01, linear learning rate warmup ($t_{warm} = 1.5K$), followed by a linear decay. During training, data augmentation is applied through random horizontal flipping, photometric distortion, and random cropping to 512×512. The model is trained for 80K iterations. λ is set to 0.1. All experiments are conducted on 4 NVIDIA GeForce RTX 3090 with PyTorch implementation.

B. Performance Comparison

Although the prevalent choice for UDA has been DeepLabV2 [2] with a ResNet-101 backbone, recent developments highlight the potential of Transformer-based SegFormer in UDA [27], [28]. Notably, while prior UDA approaches have displayed impressive results on various autonomous driving datasets, their computational demands render them unsuitable for real-time hardware configurations. To ensure a fair evaluation, we benchmark our approach against two contemporary SegFormer-based UDA methods, DAFormer [27], and HRDA [28]. For both these methods, we use the MiT-B0 encoder for training. Lastly, we conduct an exhaustive evaluation of the proposed QuadFormer on synthetic-to-real domain adaptation scenarios, specifically AutelPL Synthetic→TTPLA and AutelPL Synthetic→AutelPL Real. Extensive experiments and ablation studies substantiate the superiority of our model, with performance measured using the Intersection over Union (IoU) metric.

TABLE II: Comparision results of AutelPL Synthetic → TTPLA.

Method	Architecture	Params (M)	IoU
DAFormer [27]	MiT-B0	3.8	36.43
HRDA [28]	MiT-B0	3.8	38.36
Source-only	MiT-B0	3.8	28.93
QuadFormer	**MiT-B0**	**3.8**	**40.35**

TABLE III: Comparision results of AutelPL Synthetic → AutelPL Real.

Method	Architecture	Params (M)	IoU
DAFormer [27]	MiT-B0	3.8	35.22
HRDA [27]	MiT-B0	3.8	37.46
Source-only	MiT-B0	3.8	25.67
QuadFormer	**MiT-B0**	**3.8**	**39.65**

AutelPL Synthetic→TTPLA We first evaluate our method by utilizing AutelPL Synthetic as the source domain and TTPLA as the target domain. The performance is assessed based on the model's ability to predict pixels corresponding to the PL class on the TTPLA validation set. Our method is compared with existing state-of-the-art models by using MiT-B0 as the backbone architecture. As indicated in Table. II, the QuadFormer achieves a state-of-the-art performance of 40.35 IoU, outperforming other baselines. Compared to other transformer-based UDA models [27], [28], our method gains up to 9% IoU improvement by utilizing cross-attentive features, revealing that domain discrepancy can be reduced by considering context adaptation.

AutelPL Synthetic→AutelPL Real The domain gap between AutelPL Synthetic and AutelPL Real is greater than AutelPL Synthetic and TTPLA due to the highly diverse nature of the AutelPL Real dataset. In Table. III, we present the adaptation results on the AutelPL Real validation set, where QuadFormer exhibits considerable improvement. Our method achieves an IoU of 39.65 and outperforms all baselines by upto 6%.

C. Real-time performance

This experiment serves to demonstrate the real-time capability of our approach, making it well-suited for online operations, particularly in embedded systems. To enhance

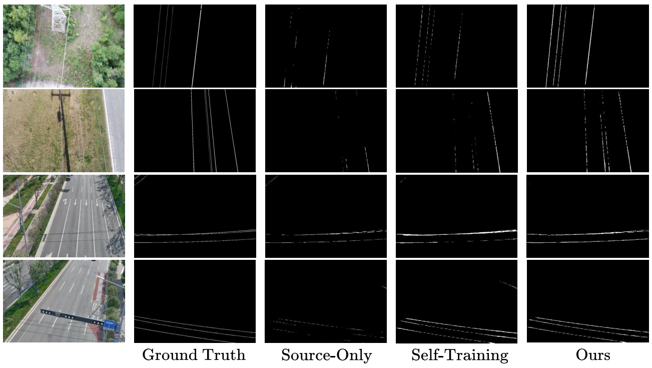

Fig. 4: QuadFormer Component Analysis comparing TTPLA (top two rows) and AutelPL Real (bottom two rows). Source-only model predictions reveal a domain gap, while self-training leveraging cross-domain features enhances prediction accuracy. Optimal results are achieved with pseudo-label correction using all three components.

inference speed, we optimized the model using TensorRT. We conducted testing on the NVIDIA Jetson NX platform, known for its compact footprint, which renders it suitable for deployment in aerial vehicles. Table IV illustrates the inference speed of the model with different image resolutions.

TABLE IV: Inference Speed on NVIDIA Jetson NX for Different Image Resolutions

Image Size	512×256	512×512	1024×512	1024×1024
Speed (Hz)	42.36	36.75	24.68	16.14

D. Ablation study

We perform extensive ablation studies to demonstrate the key components of our proposed UDA model. In the ablation studies, we train the model for 40K iterations and validated our approach on AutelPL Synthetic→AutelPL Real.

Effect of cross-attention. In Table. V, we perform ablation studies to study the effect of source and target features using only the self-attention and using both the self-attention and cross-attention. In contrast to only self-attentive features in both domains, incorporating cross-attention in the source features or target features improves the IoU by 1.72% and 2.86%, respectively. By introducing cross-attention in both domains, we achieve 39.65 IoU. While the self-attentive prototypes are able to correct the noisy pseudo labels up to some extent, the cross-attentive prototypes are less sensitive to outliers due to context adaptation. Hence, the cross-domain context at the feature level is important to understand the semantic distribution in both domains and reduces the discrepancy in data distributions.

Effect of key components of the UDA framework. Table VI indicates an ablation study of each proposed component. The source-only model gives 25.67 IoU on the target domain. Initialized by the source-only model, self-training with pseudo-labels achieves an IoU gain of 30%. Adding the adversarial loss brings a 38% IoU gain. Finally, with cross-attentive prototypes for online pseudo-label correction and all other components, our model achieves 39.65 IoU. Figure. 4 illustrates the effect of key components of the proposed QuadFormer.

TABLE V: Ablation study on the AutelPL Synthetic→AutelPL Real adaptation to understand the contribution of context adaptation.

Source features	Target features	IoU
self-attention	self-attention	34.62
cross-attention	self-attention	36.34
self-attention	cross-attention	37.48
cross-attention	cross-attention	39.65

TABLE VI: Ablation study of each proposed component on the AutelPL Synthetic→AutelPL Real adaptation.

Self Training	Adversarial Loss	Pseudo Label Correction	IoU
			25.67
✓			33.5
✓	✓		35.56
✓	✓	✓	39.65

VI. Conclusions

In this study, we introduced QuadFormer, a real-time approach tailored for unsupervised PL segmentation within the UDA framework. Our proposed solution effectively leverages information from both a labeled source domain and an unlabeled target domain, utilizing transferable context to enhance segmentation accuracy. The framework integrates cross-attention and self-attention mechanisms, enabling robust feature representations to handle domain shifts. Additionally, we incorporate self-training and pseudo-label correction mechanisms to further improve detection accuracy. To advance research in PL segmentation and evaluate our framework, we introduced two new datasets: AutelPL Synthetic and AutelPL Real. Experimental results demonstrate that QuadFormer achieves state-of-the-art performance on both AutelPL Synthetic → TTPLA and AutelPL Synthetic → AutelPL Real tasks, validating the efficacy and potential of our proposed approach for real-time PL segmentation.

References

[1] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

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

[3] ——, “Semantic image segmentation with deep convolutional nets and fully connected crfs. arxiv 2014,” *arXiv preprint arXiv:1412.7062*, 2014.

[4] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, “Segmenter: Transformer for semantic segmentation,” in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.

[5] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[6] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6881–6890.

[7] Ö. E. Yetgin, B. Benligiray, and Ö. N. Gerek, "Power line recognition from aerial images with deep learning," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 5, pp. 2241–2252, 2018.

[8] H. Zhang, W. Yang, H. Yu, H. Zhang, and G.-S. Xia, "Detecting power lines in uav images with convolutional features and structured constraints," *Remote Sensing*, vol. 11, no. 11, p. 1342, 2019.

[9] R. Abdelfattah, X. Wang, and S. Wang, "Ttpla: An aerial-image dataset for detection and segmentation of transmission towers and power lines," in *Computer Vision – ACCV 2020*, H. Ishikawa, C.-L. Liu, T. Pajdla, and J. Shi, Eds. Cham: Springer International Publishing, 2021, pp. 601–618.

[10] R. Madaan, D. Maturana, and S. Scherer, "Wire detection using synthetic data and dilated convolutional networks for unmanned aerial vehicles," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3487–3494.

[11] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2019, pp. 5982–5991.

[12] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 414–12 424.

[13] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics: Results of the 11th International Conference*. Springer, 2018, pp. 621–635.

[14] R. Kasturi and O. I. Camps, "Wire detection algorithms for navigation," 2002.

[15] J. Candamo, R. Kasturi, D. Goldgof, and S. Sarkar, "Detection of thin lines using low-quality video from low-altitude aircraft in urban settings," *IEEE Transactions on aerospace and electronic systems*, vol. 45, no. 3, pp. 937–949, 2009.

[16] C. Pan, X. Cao, and D. Wu, "Power line detection via background noise removal," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2016, pp. 871–875.

[17] J. Gubbi, A. Varghese, and P. Balamuralidhar, "A new deep learning architecture for detection of long linear infrastructure," in *Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, 2017, pp. 207–210.

[18] C. Sampedro, C. Martinez, A. Chauhan, and P. Campoy, "A supervised approach to electric tower detection and classification for power line inspection," in *International Joint Conference on Neural Networks (IJCNN)*, 2014, pp. 1970–1977.

[19] H. Choi, G. Koo, B. J. Kim, and S. W. Kim, "Weakly supervised power line detection algorithm using a recursive noisy label update with refined broken line segments," *Expert Systems with Applications*, vol. 165, p. 113895, 2021.

[20] R. Gong, W. Li, Y. Chen, and L. V. Gool, "Dlow: Domain flow for adaptation and generalization," in *IEEE/CVF Conference on Computer Vision and pattern Recognition (CVPR)*, 2019, pp. 2477–2486.

[21] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7472–7481.

[22] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *IEEE/CVF Conference on Computer Vision and pattern Recognition (CVPR)*, 2019, pp. 2517–2526.

[23] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2. Atlanta, 2013, p. 896.

[24] N. Araslanov and S. Roth, "Self-supervised augmentation consistency for adapting semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 384–15 394.

[25] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[26] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "Dacs: Domain adaptation via cross-domain mixed sampling," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WCACV)*, 2021, pp. 1379–1389.

[27] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9924–9935.

[28] ——, "Hrda: Context-aware high-resolution domain-adaptive semantic segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 372–391.

[29] X. Wang, P. Guo, and Y. Zhang, "Domain adaptation via bidirectional cross-attention transformer," *arXiv preprint arXiv:2201.05887*, 2022.

[30] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6936–6945.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.