PPTFormer: Pseudo Multi-Perspective Transformer for UAV Segmentation

Deyi Ji[1,2] , **Wenwei Jin**[2] , **Hongtao Lu**[3] and **Feng Zhao**[1*]
[1]University of Science and Technology of China
[2]Alibaba Group
[3]Dept. of CSE, MOE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
jideyi@mail.ustc.edu.cn, fzhao956@ustc.edu.cn

Abstract

The ascension of Unmanned Aerial Vehicles (UAVs) in various fields necessitates effective UAV image segmentation, which faces challenges due to the dynamic perspectives of UAV-captured images. Traditional segmentation algorithms falter as they cannot accurately mimic the complexity of UAV perspectives, and the cost of obtaining multi-perspective labeled datasets is prohibitive. To address these issues, we introduce the PPTFormer, a novel **P**seudo Multi-**P**erspective **T**rans**former** network that revolutionizes UAV image segmentation. Our approach circumvents the need for actual multi-perspective data by creating pseudo perspectives for enhanced multi-perspective learning. The PPTFormer network boasts Perspective Representation, novel Perspective Prototypes, and a specialized encoder and decoder that together achieve superior segmentation results through Pseudo Multi-Perspective Attention (PMP Attention) and fusion. Our experiments demonstrate that PPTFormer achieves state-of-the-art performance across five UAV segmentation datasets, confirming its capability to effectively simulate UAV flight perspectives and significantly advance segmentation precision. This work presents a pioneering leap in UAV scene understanding and sets a new benchmark for future developments in semantic segmentation.

1 Introduction

Semantic segmentation is a fundamental component of many practical applications in computer vision [Chen *et al.*, 2023; Yang *et al.*, 2024; Yang *et al.*, 2022a; Yang *et al.*, 2022b; Yang *et al.*, 2023; Chen *et al.*, 2024a; Yu *et al.*, 2022a; Yu *et al.*, 2024], such as autonomous driving, video surveillance, and precision agriculture, where the goal is to predict a category label for each pixel in an image. In recent years, a plethora of algorithms and networks [Zhu *et al.*, 2021; Zhu *et al.*, 2024a; Ji *et al.*, 2023c; Ji *et al.*, 2022; Zhu *et al.*, 2023c; Zhu *et al.*, 2023b] have been developed for segmentation in conventional scenes, showcasing significant advancements in the field.

With the burgeoning growth of Unmanned Aerial Vehicles (UAVs), UAV segmentation has emerged as a pivotal area of research, playing a crucial role in applications ranging from environmental monitoring to disaster response. Unlike data captured from fixed perspectives, UAVs operate at varying altitudes and angles, offering a wealth of dynamic viewpoints. The images captured by UAVs reflect the rich and varied visual information from different angles and altitudes, which is of great importance for monitoring rapidly changing environments.

Addressing the segmentation challenges posed by such rich variability in perspectives necessitates a novel approach. The most intuitive solution would be to collect actual multi-perspective data for training segmentation networks. However, as analyzed in [Ji *et al.*, 2024b; Zheng *et al.*, 2020; Yang and Ma, 2022; Ji *et al.*, 2024a], due to the prohibitive costs of collection and fine-grained annotation, existing datasets lack multi-perspective images with detailed labeling. Conventional segmentation methods typically rely on explicit perspective transformation-based data augmentation techniques, such as scaling, rotating, or flipping images in 2D or 3D dimensions. These rudimentary methods of augmentation produce stiff and unnatural perspectives that fail to represent the true changes in perspective experienced during UAV flight, resulting in limited model performance in real-world UAV scenarios.

In light of these challenges and the rapid development of Vision Transformers in semantic segmentation, a subset of methodologies [Xie *et al.*, 2021; Zheng *et al.*, 2021] has been applied to UAV scene segmentation. However, these Transformer networks have not deeply analyzed or been designed with UAV perspective dynamics in mind. The DLPL [Ji *et al.*, 2024b] introduces a comprehensive approach for latent perspective learning. However, DLPL employs a more intricate method involving discrete perspective decomposition and statistical representation, and relies on Homography transformations for converting latent perspectives. Consequently, this framework requires a slightly larger number of parameters and incurs higher latency. In this paper, we adopt the theoretical foundation of DLPL to propose a more lightweight and innovative network, termed PPTFormer (**P**seudo Multi-**P**erspective **T**rans**former**), specifically tailored for UAV seg-

*Corresponding author.

mentation tasks. The core of this network is composed of advanced PPTFormer blocks designed for pseudo multi-perspective learning. These blocks utilize perspective prototypes that remain consistent throughout the network, thereby enhancing perspective-aware learning. Drawing inspiration from [Ji *et al.*, 2024b], a critical component of the PPTFormer Blocks is the perspective representation and generation module. This module modifies visual features to emulate various UAV viewpoints while maintaining the integrity of scene semantics. Distinct from DLPL, PPTFormer utilizes a more lightweight contourlet transformation for texture extraction, followed by keypoint-based perspective description. To further reduce network complexity, we employ a non-parametric method for perspective prototype construction based on linear moving averages, coupled with a simplified Gaussian mixture distribution for generating new pseudo perspectives efficiently. We also utilize a more efficient all-MLP visual decoder instead of the complex architecture in DLPL. Moreover, to address domain shift in the multi-perspective learning process, a challenge which DLPL has overlooked, we improve the sequential attention mechanism (PIA in DLPL) with the iterative Pseudo Multi-Perspective (PMP) Attention, as well as utilize a lightweight Perspective Calibration process following the PPTFormer Blocks.

Our contributions can be summarized as follows:

- We propose a lightweight PPTFormer that enables implicit multi-perspective learning in the absence of multi-perspective datasets. By generating pseudo multi-perspective characterization about the scene and engaging in joint learning across them, PPTFormer can effectively simulate the varying viewpoints encountered during actual UAV flight, thereby improving the segmentation accuracy.
- Particularly, we employ contourlet transformation and keypoint-based perspective description, coupled with non-parametric perspective prototype construction. Moreover, a perspective calibration process is utilized to address domain shift in multi-perspective learning.
- PPTFormer achieves competitive performance across five UAV segmentation datasets while maintaining a lightweight network, demonstrating its effectiveness in capturing the intricate dynamics of UAV-captured scenes through pseudo multi-perspective Learning.

2 Related Work

2.1 Semantic Segmentation

Semantic Segmentation has been a classical and fundamental task in computer vision area [Chen *et al.*, 2024b; Ji *et al.*, 2024a; Ji *et al.*, 2020; Zhu *et al.*, 2024b; Wang *et al.*, 2021b; Wang *et al.*, 2021a; Feng *et al.*, 2018; Ji *et al.*, 2019; Ji *et al.*, 2023a; Zhu *et al.*, 2023e; Zhu *et al.*, 2023f; Zhu *et al.*, 2023d; Yu *et al.*, 2023]. In recent years, the majority of segmentation advancements are grounded in the use of fully convolutional networks (FCN) [Zhu *et al.*, 2023c; Zhu *et al.*, 2023b; Long *et al.*, 2015; Zheng *et al.*, 2023a; Zheng *et al.*, 2023b; Yu *et al.*, 2022b; Zhou *et al.*, 2023; Yu *et al.*, 2023]. Subsequent research has concentrated on capturing contextual relationships within images, employing sophisticated network architectures to enhance the understanding of scene composition [Hu *et al.*, 2020; Ji *et al.*, 2022; Zhu *et al.*, 2023a].

2.2 UAV Scene Segmentation

Despite these advancements, there is a noticeable gap in the literature pertaining to semantic segmentation tailored for UAV imagery. Existing UAV segmentation methods mainly focus on solving the class imbalanced problems, such as SCO [Yang and Ma, 2022], FarSeg [Zheng *et al.*, 2020] and PointFlow [Li *et al.*, 2021]. The unique challenges posed by UAV scenes, characterized by diverse and dynamic changes in perspective, are seldom addressed. The development of segmentation models that can effectively handle the variability inherent in UAV-captured images also remains an area in need of further exploration.

2.3 Ultra-High Resolution Segmentation

In comparison to natural scene imagery, images captured from Unmanned Aerial Vehicles (UAVs) generally exhibit higher resolution characteristics. Existing research has also attempted to introduce segmentation algorithms that can concurrently balance accuracy and efficiency, such as WSDNet [Ji *et al.*, 2023c], GPWFormer [Ji *et al.*, 2023b], among others. In this paper, we endeavor to enhance segmentation precision from the perspective of UAV flight viewpoints. This approach is universally applicable and can be integrated with the aforementioned high-resolution image segmentation algorithms.

3 PPTFormer

3.1 Overall Structure

The overall architecture of our proposed Pseudo Multi-Perspective Transformer (PPTFormer) is inspired by [Ji *et al.*, 2024b] and depicted in Figure 1. Following the classic paradigms [Zheng *et al.*, 2021; Xie *et al.*, 2021], PPTFormer consists of a meticulously designed encoder for perspective learning and a generic decoder. The encoder comprises four Transformer Blocks: one Plain Transformer Block followed by three PPTFormer Blocks. The former is responsible for extracting basic low-level information, which serves as the foundation for pseudo multi-perspective learning in the subsequent blocks. Unlike [Ji *et al.*, 2024b] which applies DLPL at every Transformer Block, we restrict perspective learning to only the last three Blocks in PPTFormer to enhance its efficiency. Specifically, within the PPTFormer Blocks, we extract implicit perspective representations from the visual features based on the structural texture. In conjunction with training across the entire dataset, we create perspective prototypes of the images present throughout the dataset. These prototypes are shared across the three PPTFormer Blocks to ensure a consistent learning of perspectives during the training process. Interleaved between the PPTFormer Blocks are Perspective Calibration modules, which are instrumental in aligning the visually fused features from pseudo perspectives with the original perspective of the image. This alignment prevents potential scene domain shifts, which are overlooked

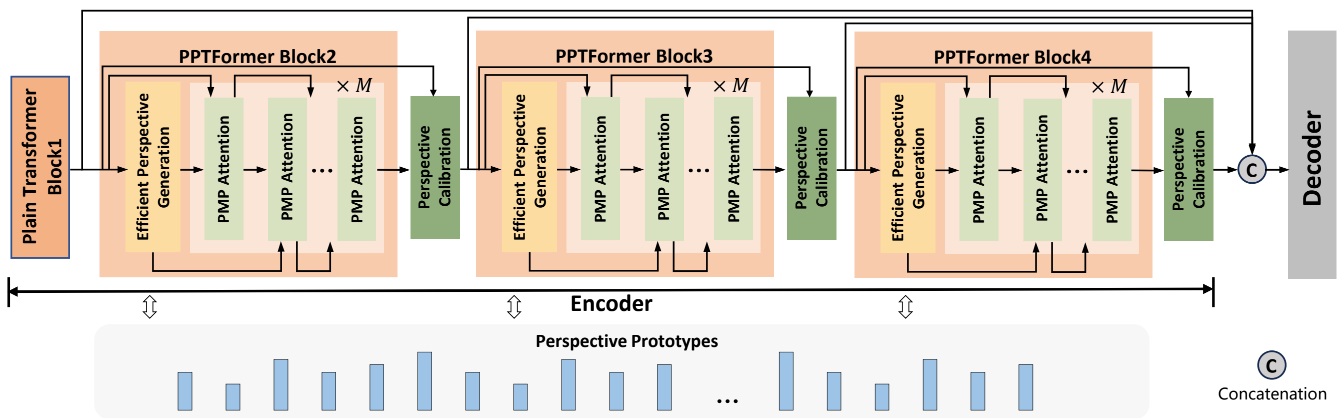

Figure 1: The overview of the proposed PPTFormer. The encoder comprises four Transformer Blocks: one Plain Transformer Block followed by three PPTFormer Blocks. The former is responsible for extracting basic low-level information, which serves as the foundation for pseudo multi-perspective learning in the subsequent blocks. Interleaved between the PPTFormer Blocks are Perspective Calibration modules, which prevent potential scene domain shifts. Finally, we concatenate features of varying scales produced by each block and feed them into the decoder network.

in DLPL [Ji *et al.*, 2024b]. Finally, we concatenate features of varying scales produced by each block and feed them into the decoder network for further processing.

3.2 PPTFormer Block

As shown in Figure 1, the PPTFormer Block comprises an Efficient Perspective Generation module and M layers of Pseudo Multi-Perspective Attention (PMP Attention). Given the input of low-level visual features F from block 1, the Efficient Perspective Generation module implicitly represents the image's perspective p, generating a pseudo perspective p' that simulates the movement and shift of viewpoints during an actual UAV flight, all while preserving the semantic information of the scene. During this process, the acquired perspective representation p contributes to the construction of perspective prototypes P for the entire dataset and also bases the pseudo perspective generation on these prototypes. The output visual feature F' with the pseudo perspective p', along with F, are both fed into the M layers of PMP Attention for multi-perspective correlation. This allows the model to understand the scene's semantic information from both the original perspective and the new pseudo perspective simultaneously. Specifically, in the first layer of PMP Attention, the inputs are F and F', and the output is the first level of perspective fusion. Subsequently, in the following $M-1$ layers of PMP Attention, the fused feature and F' work together to achieve the subsequent $M-1$ levels of perspective fusion. Below, we will introduce the specific structures of the Efficient Perspective Generation and PMP Attention in detail.

3.3 Efficient Perspective Generation

As illustrated in Figure 2, the input is the visual feature F form Block 1, which first passes through a Perspective Representation encoder $\mathbf{E}_p$ to obtain an original Perspective p. Subsequently, on one hand, p contributes to the construction of the entire dataset's perspective prototypes P using the online sequential clustering updating technique. The length of P, which corresponds to the number of perspective prototypes in the dataset, is N. On the other hand, p is also used for Pseudo Perspective Generation based on P, to generate a new pseudo perspective p'. Thereafter, a Visual Reconstruction decoder $\mathbf{D}_p$ uses p' to ultimately reconstruct the visual feature F'. In order to reduce the complexity, $\mathbf{D}_p$ is all-MLP architecture which demonstrates promising performance here. During training, to ensure its reconstructive capability, p is also directly fed into $\mathbf{D}_p$ with the aim of restoring the original visual feature F, that is,

$$L_{rec} = ||\mathbf{D}_p(p) - F||_2 = ||\mathbf{D}_p(\mathbf{E}_p(F)) - F||_2, \quad (1)$$

where L_{rec} is the reconstruction loss.

Perspective Representation

As depicted in Figure 3, distinct from DLPL, PPTFormer utilizes a more lightweight contourlet transformation for texture extraction, followed by keypoint-based perspective description. The Perspective Representation encoder $\mathbf{E}_p$ primarily encompasses two processes: extracting low-level structural texture from the image using contourlet decomposition [Do and Vetterli, 2005] and, based on this texture, extracting interest super points that are related to the image's perspective. The former ensures that the model captures the global structural texture information, which can represent the image's contours, edges, and other structural features, including perspective information. To further distill perspective-related features, the latter identifies key support points representing perspective as super points, whose spatial distribution and feature intensity can construct a structured high-dimensional description of the image's perspective.

Texture Extraction. Specifically, as traditional filters, contourlet decompositions inherently excel at texture representation across various geometric scales and directions in the spectral domain. Rather than describing texture features in the spatial domain, they analyze the energy distribution in the spectral domain to extract the inherent geometric structures of the texture, which naturally includes the image's perspective.

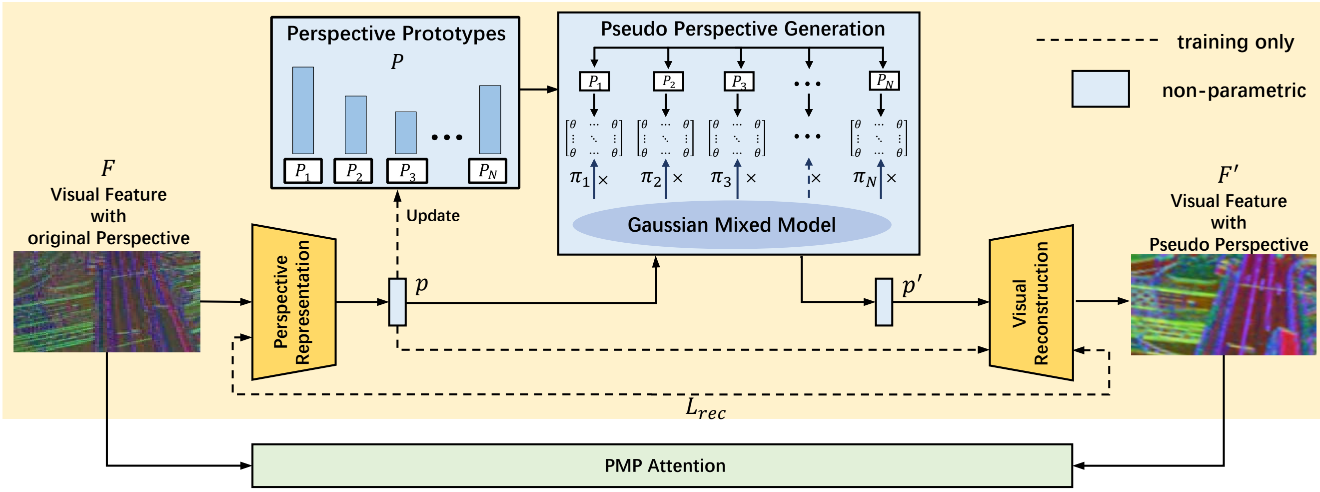

Figure 2: The Efficient Perspective Generation module in PPTFormer Block.

The contourlet decomposition comprises a cascaded Laplacian Pyramid (LP) [Burt and Adelson, 1983] and a directional filter bank (DFB) [Bamberger and Smith, 1992]. The LP decomposes input features into low-pass and high-pass subbands using pyramidal filters. The high-pass subband is processed through the DFB, which is employed to reconstruct the original signal with minimal sample representation, produced by t-level binary tree decomposition in the two-dimensional frequency domain, resulting in 2^z directional subbands. For instance, when $z = 3$, the frequency domain is divided into 8 directional subbands, with subbands 0-3 and 4-7 corresponding to vertical and horizontal details, respectively [Ji *et al.*, 2022]. Following [Ji *et al.*, 2022], for a richer expression, we stack multiple contourlet decomposition layers iteratively, and concatenate the output of each level to form the final extracted structural texture.

Specifically, the output of level $t \in [1, T]$ is denoted as $F_{bds,t}$,

$$F_{bds,t} = \mathbf{DFB}(F_{h,t}), \quad t \in [1, T], \\ \text{where } F_{l,t}, F_{h,t} = \mathbf{LP}(F_{l,t-1}) \tag{2}$$

where l and h represent the low-pass and high-pass subbands respectively, bds denotes the bandpass directional subbands. Then structural texture $F_{texture}$ is denoted as:

$$F_{texture} = \underset{t\in[1,T]}{\mathrm{Cat}} \{F_{bds,t}\}. \tag{3}$$

where Cat is the concatenation operation. $F_{texture}$ is rich in texture information including the image's perspective.

Perspective Support Description. Based on $F_{texture}$, we use a SuperPoint network to extract key support points and corresponding support descriptors that are capable of characterizing the perspective from the texture features. Specifically, this network comprises two parallel heads, $\mathbf{S}_{sp}$ and $\mathbf{S}_{sd}$, which respectively output the "point-ness" probability map and the corresponding point feature descriptor. The final output, the perspective feature p, is the concatenation of output features from the two heads, along the channel dimension:

$$p = \mathrm{Cat}(\mathbf{S}_{sp}(F_{texture}), \mathbf{S}_{sd}(F_{texture})). \tag{4}$$

Perspective Prototypes Construction

The construction of Perspective Prototypes is aimed to obtain and manage the scene perspective types of the whole dataset, by performing an online sequential clustering process on the coming ps. We utilize a lightweight memory bank and its length N is equal to the number of prototypes. Specifically, different from DLPL [Ji *et al.*, 2024b], we employ a non-parametric mechanism to reduce the network complexity. Firstly, the N prototypes $P = \{P_1, P_2, ..., P_n, ..., P_N\}$ are initialized with the first input N ps, and we set the counts $\{c_1, c_2, ..., c_n, ..., c_N\}$ to record the number of perspective features belonging to the corresponding prototype. Then, for each new coming p, we find its closest prototype P_n by L2 distances, and update the the prototype with:

$$c_n \leftarrow c_n + 1 \\ P_n \leftarrow P_n + \frac{1}{c_n}(p - P_n). \tag{5}$$

So the final resulting prototype P_n is the moving average of ps that are closest to P_n.

Then we formulate overall perspective distribution of the whole dataset in form of Gaussian Mixed Model (GMM), following DLPL,

$$G(P) = \sum_{n=1}^{N} \pi_n \cdot \mathcal{N}(p|P_n, \Sigma_n), \tag{6}$$

where $\mathcal{N}(\cdot)$ indicates the Gaussian Distribution, the nth component of GMM has the center of P_n with the variance of Σ_n, and π_n is the mixture coefficient of meets:

$$\sum_{n=1}^{N} \pi_n = 1, \quad 0 \le \pi_n \le 1, \tag{7}$$

Pseudo Perspective Generation

Based on the dynamically updated perspective distribution $G(P)$, we can generate a new semantic-related pseudo perspective p' of the given probe p, by leveraging all the proto-

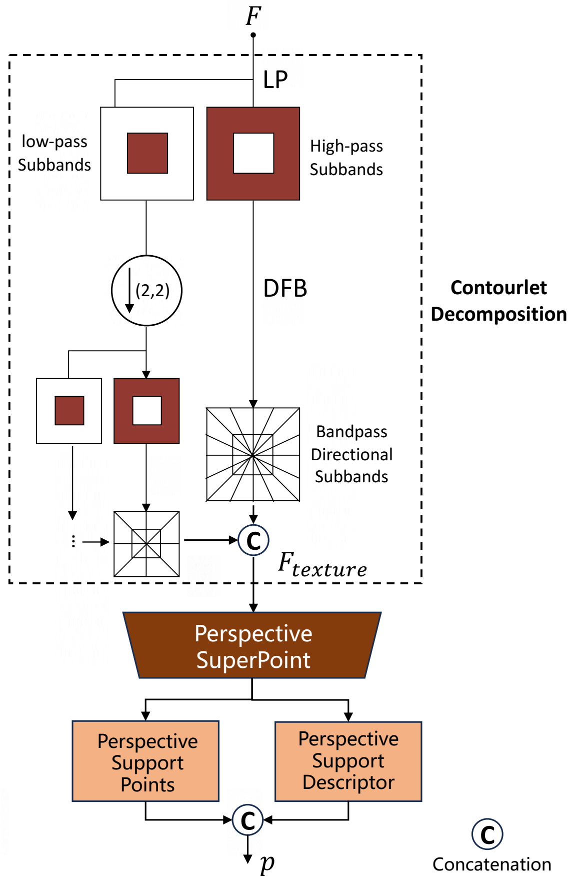

Figure 3: The detailed process of Perspective Representation.

types. Note that we avoid using Homography transformations in DLPL to accelerate the generating process.

$$p' = p \cdot G(P, p) = p \cdot \sum_{n=1}^{N} \pi_n \cdot \mathcal{N}(p|P_n, \Sigma_n), \tag{8}$$

where p' is generated based on the overall perspective distribution over all perspective prototypes.

Finally, the corresponding visual feature F' for p' can be reconstructed with $\mathbf{D}_p$.

$$F' = \mathbf{D}_p(p'). \tag{9}$$

3.4 Pseudo Multi-Perspective Attention

By the Efficient Perspective Generation, we obtain the visual feature F' with a new pseudo perspective for the input image. As seen that F and F' contain closely identical scene context and structured information but only differ in perspective. Next, they are fed into the M layers of Pseudo Multi-Perspective (PMP) Attention to leverage the relationship between the F (with original perspective p) and F' (with generated pseudo perspective p'). Different from DLPL which use sequential attention mechanism, the PMP Attention employs an iterative one,

$$\text{PMP\_Att}(F, F') = \begin{cases} F_1 = \text{Softmax}(\frac{F \times F'^{\top}}{\sqrt{C}}) \times F', \\ F_2 = \text{Softmax}(\frac{F' \times F_1^{\top}}{\sqrt{C}}) \times F_1, \\ F_3 = \text{Softmax}(\frac{F_1 \times F_2^{\top}}{\sqrt{C}}) \times F_2, \\ \dots \\ F_M = \text{Softmax}(\frac{F_{M-2} \times F_{M-1}^{\top}}{\sqrt{C}}) \times F_{M-1}. \end{cases}, \tag{10}$$

where C is the feature channel. By the iterative attention calculation, the multiple perspectives can be fully correlated.

3.5 Perspective Calibration

Within each PPTFormer Block, after undergoing N layers of PMP Attention, the original perspective and the pseudo perspective are thoroughly fused multiple times, ultimately enabling the model to capture scene information as observed from various perspectives. However, in practice, we observed that as the perspective fusion progresses, there could be some domain shift within the visual features' depiction of the scene. To prevent such occurrences, we further incorporate a straightforward Perspective Calibration process after PPTFormer Blocks. Specifically, this entails passing the visual feature with the original perspective, which is the input to the current PPTFormer Block, through a skip connection to calibrate the fused feature output by the current PPTFormer Block, by several layers of PMP Attention. In practice, we found this elegant approach to be effective in mitigating issues of domain shift.

3.6 Optimization

The overall loss function L is the combination of the main segmentation loss L_{seg} and the reconstruction loss L_{rec}:

$$L = L_{seg} + \lambda L_{rec}, \tag{11}$$

where λ is the weight for L_{rec}, and set to 0.4.

4 Experiments

4.1 Datasets and Evaluation Metrics

In our experiments, we validate the effectiveness of PPTFormer on five datasets, including UDD6, iSAID, UAVid , Aeroscapes, and DroneSeg.

UDD6

Urban Drone Dataset (UDD) dataset is collected by a DJI-Phantom 4 UAV at altitudes between 60 and 100 meters, and is extracted from 10 video sequences. The resolution is either 4k (4096×2160) or 12M (4000×3000). It contains a variety of urban scenes.

iSAID

iSAID totally consists of 2,806 images, where 1411, 458, and 937 images are for training, validation, and testing sets, respectively.

Method	mIoU (%)				
	UDD6	iSAID	UAVid	Aeroscapes	DroneSeg
Deeplab	71.84	59.20	56.82	51.40	38.69
OCR_W48	73.37	62.73	63.10	58.19	43.10
PSPNet	72.95	60.30	58.20	57.98	37.03
FarSeg	-	63.70	-	-	-
FarSeg++	-	63.70	-	-	-
PFNet	-	66.90	-	-	-
SCO	-	69.10	-	-	-
SETR	68.00	62.77	58.52	50.34	48.23
UperNet	73.13	66.45	61.91	64.32	53.34
PoolFormer	74.54	65.55	61.73	62.27	53.94
SegFormer	74.28	67.19	62.01	66.40	55.33
PPTFormer	**76.70**	**69.87**	**65.00**	**68.50**	**57.71**

Table 1: Comparison with state-of-the-arts on UDD6, iSAID, UAVid, Aeroscapes, and our proposed DroneSeg datasets.

UAVid

UAVid dataset has 300 images of size of 3840×2160, where the training, validation, and testing set contains 200, 70, and 30 images respectively.

Aeroscapes

The Aeroscapes dataset provides 3,269 720p images and ground-truth masks for 11 categories, where the training and validation sets include 2,621 and 648 images respectively.

DroneSeg

The DroneSeg dataset [Ji *et al.*, 2024b] extends the segmentation annotations from VisDrone dataset. The dataset consists of 10,209 images with fine-grained pixel-level annotations of 14 categories.

4.2 Implementation Details

In our experiments, we adopt the MMSegmentation toolbox as codebase and follow the default augments without bells and whistles. SuperPoint network is used for the perspective support description. To ensure training stability, during the initial 30% of epochs, we replace the PMP Attention with plain Self-Attention. This substitution aims to guarantee the reliability of perspective representation and reconstruction within $\mathbf{E}_p$ and $\mathbf{D}_p$, as well as the stability of learning perspective prototypes. Subsequently, we revert to the PMP Attention mechanism to perform global joint optimization in the remaining epochs. In the training, SGD optimizer with momentum 0.98 for all parameters is used, the initial learning rate is configured as 5×10^{-3} and the maximum iteration number is set to 160K for all datasets. In Eq. 2, T is set to 2. The length of P is set to $N = 64$.

4.3 Comparison with State-of-the-Arts

We compare PPTFormer with both representative CNN-based (FarSeg [Zheng *et al.*, 2020], FarSeg++ [Zheng *et al.*, 2020], PFNet [Li *et al.*, 2021] and SCO [Yang and Ma, 2022]) and ViT-based (SETR [Zheng *et al.*, 2021], UperNet [Liu *et al.*, 2021], PoolFormer [Yu *et al.*, 2022c], SegFormer [Xie *et al.*, 2021]) segmentation methods on five benchmark datasets.

Perspective-Oriented Learning Method	mIoU (%)
Baseline (SegFormer w/o data aug.)	52.03
+ Random Rotate	52.94
+ Random Scale	53.09
+ Random Perspective-Vertical	53.56
+ Random Perspective-Horizontal	53.42
+ Random Combination	55.33
PPTFormer	57.71

Table 2: The comparison of PPTFormer with other perspective-oriented learning methods (data augmentation). Perspective-Vertical and Perspective-Horizontal means adjusting perspectives in vertical and horizontal directions, and are implemented by the default "torchvision.transforms" interfaces in PyTorch.

UDD6, UAVid

Both the two datasets contain relative fewer images and lower scene complexity and we compare their results here. For fair comparisons with ViT-based methods, we adopt large backbones for CNN-based methods including ResNet-101 and HRNet-W48. As shown in Table 1, the ordinary transformer (SETR) show even lower performance than CNN-based methods, and the advanced ones including PoolFormer, SegFormer shows better results. The proposed PPTFormer achieves further performance improvements on both the datasets.

iSAID, Aeroscapes, DroneSeg

These three datasets consist of more images than UDD6 and UAVid, and have higher scene perspective variances. So they would be more convincing to prove the superiority. PPTFormer outperforms other methods by a larger margin, which demonstrates the effectiveness of the proposed method on the description of perspective information.

4.4 Ablation Study

All ablation studies are performed on DroneSeg *testing* set, SegFormer is used as baseline network.

Comparison with Perspective Learning Methods

Given that PPTFormer is a perspective-oriented learning approach, we begin by comparing it with various perspective-based augmentations. We observe that in UAV scenarios, perspective shifts are almost invariably linked to changes in altitude and angular positioning, which manifest as alterations in scale and rotation. Therefore, we employ a combination of these two data augmentation techniques to benchmark against PPTFormer. As illustrated in Table 2, we discover that utilizing either augmentation method in isolation yields only modest enhancements over the baseline approach. In contrast, PPTFormer secures a substantial increase in performance. We further demonstrate that PPTFormer remains compatible with standard data augmentations for additional gains.

The Impact of Contourlet Decomposition

The contourlet decomposition is capable of extracting structural texture information from images, which encompasses a wealth of perspective details. By initially employing it within

Contourlet Decomposition	mIoU (%)
0	56.88
1	57.40
2	57.71
3	57.73

Table 3: The impact of contourlet decomposition.

Method	Parameters
SegFormer	84.7M
DLPL(Segformer)	87.5M
PPTFormer	86.0M

Table 4: Efficiency analysis.

the Perspective Representation, the network can swiftly focus on shallow image textures, thereby facilitating further extraction of Perspective and enhancing learning efficiency. Table 3 demonstrates its efficacy, as the number of contourlet decomposition layers increases, the mIoU correspondingly improves. Here, a layer count of zero indicates no application of contourlet decomposition.

Effectiveness of Perspective Calibration

The purpose of Perspective Calibration is to prevent the occurrence of scene domain shift that may arise as a consequence of deep perspective fusion. Figure 4 illustrates the impact of the number of PMP Attention layers in it on model performance. “Layer=0” implies the absence of Calibration, and the results indicate a low mIoU under this condition. As the number of layers increases, there is a significant improvement in mIoU, which underscores the effectiveness and necessity of Perspective Calibration.

The Quantity of Perspective Prototypes

The quantity of perspective prototypes represents the entirety of perspective variations found within the dataset, with a higher count enabling the retention of a more extensive set of prototypes. Figure 5 reveals that with a smaller allocation of prototypes (16, 32), the process fails to exhaustively capture all perspectives, resulting in an underfitting of the model. Conversely, as the number of prototypes increases (128, 256), we generate an overly dense array of prototypes. This surplus can introduce redundancy and give rise to numerous discrete perspectives, potentially hindering the learning process.

Efficiency Analysis

PPTFormer, inspired by the theoretical framework of DLPL, exhibits a more lightweight architecture and enhanced efficiency. In Table 4, we present a comparative analysis of the parameter counts between PPTFormer and DLPL (based on SegFormer). The comparison indicates that PPTFormer introduces only a slight increase in parameters over the baseline SegFormer, yet maintains a lower parameter count than DLPL. This demonstrates that PPTFormer not only achieves high accuracy but also sustains a high level of efficiency.

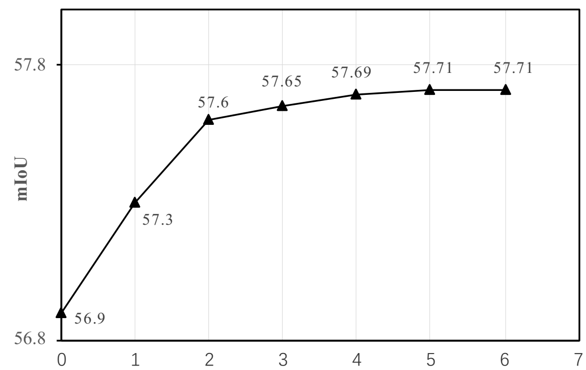

Figure 4: Layer number in perspective calibration.

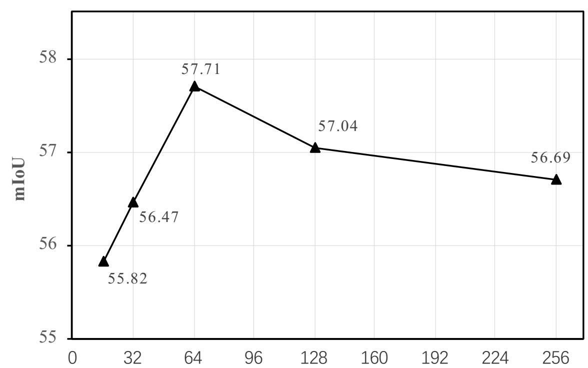

Figure 5: The quantity of perspective prototypes.

5 Conclusion

This paper presents the novel PPTFormer, a Pseudo Multi-Perspective Transformer network for UAV scene segmentation. It addresses the challenges of capturing the dynamic perspectives inherent in UAV-captured imagery. By integrating systematic Pseudo Multi-Perspective Learning within the Transformer framework, PPTFormer adeptly performs efficient pseudo perspective generation and achieves multi-perspective correlation, leading to a more nuanced understanding of UAV scenes. The experiments on several datasets validate the superior performance and efficiency of PPTFormer. The significant advancements made by PPTFormer underscore the importance of perspective-oriented learning in semantic segmentation and pave the way for further innovation in the processing of UAV-captured visual data.

Acknowledgments

This work was supported by the Anhui Provincial Natural Science Foundation under Grant 2108085UD12, the National Key R&D Program of China under Grant 2020AAA0103902, NSFC (No. 62176155), Shanghai Municipal Science and Technology Major Project, China (2021SHZDZX0102). We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

Contribution Statement

The first two authors contribute equally to this work.

References

[Bamberger and Smith, 1992] R.H. Bamberger and M.J.T. Smith. A filter bank for the directional decomposition of images: theory and design. *IEEE Transactions on Signal Processing*, 40(4):882–893, 1992.

[Burt and Adelson, 1983] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

[Chen *et al.*, 2023] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 2023.

[Chen *et al.*, 2024a] Tianrun Chen, Chaotao Ding, Lanyun Zhu, Tao Xu, Deyi Ji, Ying Zang, and Zejian Li. xlstm-unet can be an effective 2d\& 3d medical image segmentation backbone with vision-lstm (vil) better than its mamba counterpart. *arXiv preprint arXiv:2407.01530*, 2024.

[Chen *et al.*, 2024b] Tianrun Chen, Chunan Yu, Jing Li, Jianqi Zhang, Lanyun Zhu, Deyi Ji, Yong Zhang, Ying Zang, Zejian Li, and Lingyun Sun. Reasoning3d – grounding and reasoning in 3d: Fine-grained zero-shot open-vocabulary 3d reasoning part segmentation via large vision-language models. *arXiv preprint arXiv:2405.19326*, 2024.

[Do and Vetterli, 2005] M.N. Do and M. Vetterli. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12):2091–2106, 2005.

[Feng *et al.*, 2018] Weitao Feng, Deyi Ji, Yiru Wang, Shuorong Chang, Hansheng Ren, and Weihao Gan. Challenges on large scale surveillance video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 69–76, 2018.

[Hu *et al.*, 2020] Hanzhe Hu, Deyi Ji, Weihao Gan, Shuai Bai, Wei Wu, and Junjie Yan. Class-wise dynamic graph convolution for semantic segmentation. In *European Conference on Computer Vision*, pages 1–17, 2020.

[Ji *et al.*, 2019] Deyi Ji, Hongtao Lu, and Tongzhen Zhang. End to end multi-scale convolutional neural network for crowd counting. In *Eleventh International Conference on Machine Vision*, volume 11041, pages 761–766, 2019.

[Ji *et al.*, 2020] Deyi Ji, Haoran Wang, Hanzhe Hu, Weihao Gan, Wei Wu, and Junjie Yan. Context-aware graph convolution network for target re-identification. *arXiv preprint arXiv:2012.04298*, 2020.

[Ji *et al.*, 2022] Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, and Hongtao Lu. Structural and statistical texture knowledge distillation for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16876–16885, 2022.

[Ji *et al.*, 2023a] Deyi Ji, Siqi Gao, Mingyuan Tao, Hongtao Lu, and Feng Zhao. Changenet: Multi-temporal asymmetric change detection dataset. *arXiv preprint arXiv:2312.17428*, 2023.

[Ji *et al.*, 2023b] Deyi Ji, Feng Zhao, and Hongtao Lu. Guided patch-grouping wavelet transformer with spatial congruence for ultra-high resolution segmentation. *International Joint Conference on Artificial Intelligence*, pages 920–928, 2023.

[Ji *et al.*, 2023c] Deyi Ji, Feng Zhao, Hongtao Lu, Mingyuan Tao, and Jieping Ye. Ultra-high resolution segmentation with ultra-rich context: A novel benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23621–23630, June 2023.

[Ji *et al.*, 2024a] Deyi Ji, Siqi Gao, Lanyun Zhu, Qi Zhu, Yiru Zhao, Peng Xu, Hongtao Lu, Feng Zhao, and Jieping Ye. View-centric multi-object tracking with homographic matching in moving uav. *arXiv preprint arXiv:2403.10830*, 2024.

[Ji *et al.*, 2024b] Deyi Ji, Feng Zhao, Lanyun Zhu, Wenwei Jin, Hongtao Lu, and Jieping Ye. Discrete latent perspective learning for segmentation and detection. *International Conference on Machine Learning*, 2024.

[Li *et al.*, 2021] Xiangtai Li, Hao He, Xia Li, Duo Li, Guangliang Cheng, Jianping Shi, Lubin Weng, Yunhai Tong, and Zhouchen Lin. Pointflow: Flowing semantics through points for aerial image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4217–4226, 2021.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[Wang *et al.*, 2021a] Haoran Wang, Licheng Jiao, Fang Liu, Lingling Li, Xu Liu, Deyi Ji, and Weihao Gan. Ipgn: Interactiveness proposal graph network for human-object interaction detection. *IEEE Transactions on Image Processing*, 30:6583–6593, 2021.

[Wang *et al.*, 2021b] Haoran Wang, Licheng Jiao, Fang Liu, Lingling Li, Xu Liu, Deyi Ji, and Weihao Gan. Learning social spatio-temporal relation graph in the wild and a video benchmark. *IEEE Transactions on Neural Networks and Learning Systems*, 34(6):2951–2964, 2021.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *In 35th Conference on Neural Information Processing Systems*, pages 1–13, 2021.

[Yang and Ma, 2022] Fengyu Yang and Chenyang Ma. Sparse and complete latent organization for geospatial semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1809–1818, 2022.

[Yang *et al.*, 2022a] Zizheng Yang, Xin Jin, Kecheng Zheng, and Feng Zhao. Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14298–14307, 2022.

[Yang *et al.*, 2022b] Zizheng Yang, Mingde Yao, Jie Huang, Man Zhou, and Feng Zhao. Sir-former: Stereo image restoration using transformer. In *ACM international conference on multimedia*, pages 6377–6385, 2022.

[Yang *et al.*, 2023] Zizheng Yang, Jie Huang, Jiahao Chang, Man Zhou, Hu Yu, Jinghao Zhang, and Feng Zhao. Visual recognition-driven image restoration for multiple degradation with intrinsic semantics recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14059–14070, 2023.

[Yang *et al.*, 2024] Zizheng Yang, Jie Huang, Man Zhou, Naishan Zheng, and Feng Zhao. IRVR: A general image restoration framework for visual recognition. *IEEE Transactions on Multimedia*, pages 7012–7026, 2024.

[Yu *et al.*, 2022a] Hu Yu, Jie Huang, Feng Zhao, Jinwei Gu, Chen Change Loy, Deyu Meng, Chongyi Li, et al. Deep fourier up-sampling. *Advances in Neural Information Processing Systems*, 35:22995–23008, 2022.

[Yu *et al.*, 2022b] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *European Conference on Computer Vision*, pages 181–198. Springer, 2022.

[Yu *et al.*, 2022c] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10819–10829, 2022.

[Yu *et al.*, 2023] Wei Yu, Qi Zhu, Naishan Zheng, Jie Huang, Man Zhou, and Feng Zhao. Learning non-uniform-sampling for ultra-high-definition image enhancement. In *ACM International Conference on Multimedia*, pages 1412–1421, 2023.

[Yu *et al.*, 2024] Hu Yu, Jie Huang, Lingzhi Li, Feng Zhao, et al. Deep fractional fourier transform. *Advances in Neural Information Processing Systems*, 2024.

[Zheng *et al.*, 2020] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4096–4105, 2020.

[Zheng *et al.*, 2021] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2021.

[Zheng *et al.*, 2023a] Naishan Zheng, Jie Huang, Feng Zhao, Xueyang Fu, and Feng Wu. Unsupervised underexposed image enhancement via self-illuminated and perceptual guidance. *IEEE Transactions on Multimedia*, 2023.

[Zheng *et al.*, 2023b] Naishan Zheng, Jie Huang, Man Zhou, Zizheng Yang, Qi Zhu, and Feng Zhao. Learning semantic degradation-aware guidance for recognition-driven unsupervised low-light image enhancement. In *AAAI Conference on Artificial Intelligence*, pages 3678–3686, 2023.

[Zhou *et al.*, 2023] Man Zhou, Naishan Zheng, Yuan Xu, Chun-Le Guo, and Chongyi Li. Training your image restoration network better with random weight network as optimization function. In *37th Advances in Neural Information Processing Systems*, pages 1270–1282, 2023.

[Zhu *et al.*, 2021] Lanyun Zhu, Deyi Ji, Shiping Zhu, Weihao Gan, Wei Wu, and Junjie Yan. Learning statistical texture for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12537–12546, 2021.

[Zhu *et al.*, 2023a] Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llafs: When large-language models meet few-shot segmentation. *arXiv preprint arXiv:2311.16926*, 2023.

[Zhu *et al.*, 2023b] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Continual semantic segmentation with automatic memory sample selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3082–3092, 2023.

[Zhu *et al.*, 2023c] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Learning gabor texture features for fine-grained recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 1621–1631, 2023.

[Zhu *et al.*, 2023d] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Learning gabor texture features for fine-grained recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 1621–1631, 2023.

[Zhu *et al.*, 2023e] Qi Zhu, Jie Huang, Naishan Zheng, Hongzhi Gao, Chongyi Li, Yuan Xu, Feng Zhao, et al. Fouridown: Factoring down-sampling into shuffling and superposing. In *Advances in Neural Information Processing Systems*, volume 36, pages 1–14, 2023.

[Zhu *et al.*, 2023f] Qi Zhu, Man Zhou, Naishan Zheng, Chongyi Li, Jie Huang, and Feng Zhao. Exploring temporal frequency spectrum in deep video deblurring. In *IEEE/CVF International Conference on Computer Vision*, pages 12428–12437, 2023.

[Zhu *et al.*, 2024a] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Addressing background context bias in few-shot segmentation through iterative modulation. In *IEEE/CVF International Conference on Computer Vision*, pages 1–10, 2024.

[Zhu *et al.*, 2024b] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024.