STCMR-AVSS:Spatio-Temporal Context modeling and refining for Airport Video Semantic Segmentation

Caihua Liu[1,2], Wei Tang[1,2], Junhao Chen[1,2], Xia Feng[1,3],
[1]Key Laboratory of Civil Aviation Information Technology Application Innovation of Tianjin, China
[2]College of Computer Science and Artificial Intelligence, Civil Aviation University of China, Tianjin, China
[3]Science and Technology Innovation Research Institute, Civil Aviation University of China, Tianjin

***Abstract*—Existing airport video semantic segmentation methods exhibit significant limitations in spatio-temporal context modeling. The extracted motion features often lack scale adaptability, and inter-frame temporal consistency is not fully maintained. To effectively capture multi-scale motion features and fully leverage inter-frame information consistency, we propose a Spatio-Temporal Context Modeling and Refining method. Firstly, a Multi-scale Motion Dynamic Aggregation (MMDA) module is designed to flexibly learn the connections between temporal dynamics and multi-scale spatial information. It employs a joint Spatio-Temporal Context modeling strategy that not only integrates motion information across time dimensions but also incorporates multi-scale motion details. This effectively enhances the coherence and correlation between temporal dynamics and spatial features. Secondly, a Cross-frame Refinement (CFR) module is introduced to mine useful information from different frames, keeping inter-frame temporal consistency.By leveraging the synergistic effort of these two modules, we obtain comprehensive multi-scale spatial and temporal information to enhance the representation of the target frame. Finally, enhanced features are concatenated with original target frame features and fed together into a lightweight decoder for segmentation.Extensive experiments on the AVSS benchmark demonstrate significant performance gains across multiple metrics over state-of-the-art methods.**

***Index Terms*—Airport video semantic segmentation, Spatio-Temporal Context modeling, Multi-scale Motion Dynamic Aggregation, Cross-frame Refinement**

I. Introduction

Airport video semantic segmentation aims to classify each pixel in airport videos into specific categories (such as airplanes, vehicles,, etc.), while ensuring consistent segmentation results for the same object across different frames. This technology can be widely deployed for object localization and behavior analysis on the airport surface, which is vital to operational safety.

Previous video semantic segmentation research has primarily focused on image segmentation [1]–[7], exploring methods to model multi-scale features in images to enhance the capability of handling objects at various scales. In video semantic segmentation studies, current approaches mainly utilize recurrent units [8], [9], attention-based propagation mechanisms [10], [11], optical flow motion [12], [13], temporal consistency regularization [14]–[17], or spatio-temporal Transformers [18] to model or leverage temporal consistency. However, these methods largely overlook the integration of multi-scale features internal to the temporal modeling process. Recent studies [10], [19]–[21] have attempted to incorporate multi-scale features, but they typically treat temporal modeling and multi-scale feature extraction as independent processes. This decoupled approach prevents effective correlation of motion trajectories and morphological evolution across scales, leading to the loss of critical multi-scale motion context.

(a)An increased scale disparity due to temporal changes.

(b)Severe occlusion is caused by the movement.

Fig. 1. Some challenging video samples captured in airports.

However, despite these advancements, previous methods still face challenges when pursuing airport video semantic segmentation:**(a) significant scale variations over time**. In complex scenes(Fig 1(a)), objects (such as aircraft) undergo drastic scale variations synchronized over time. The decoupled approaches adopted by existing methods hinder the simultaneous capture of multi-scale features and long-term temporal dependencies, making it difficult to obtain effective spatiotemporal context features.**(b) Severe Occlusion**: In cluttered airport environments(Fig 1(b)), interactions among aircraft, pedestrians, and ground vehicles lead to severe occlusion. Most existing methods simply aggregate features

This work was supported by the Open Project of Tianjin Key Laboratory of Civil Aviation Information Technology Application Innovation (Grant No. CAITAIT-202403); the 2024 Tianjin "the Belt and Road" Joint Laboratory Project under the Special Innovation Platform Program of Tianjin Science and Technology Plan (China Kazakhstan Joint Research Center for Digital Twin Airport, Civil Aviation University of China, Grant No. 24PTLYHZ00230); and the Civil Aviation Joint Fund of the National Natural Science Foundation of China (Grant No. U2333206).

across different frames but pay little attention to mining and refining the consistency of spatiotemporal information between frames. This makes it difficult to obtain highly consistent inter-frame semantic information in a complementary manner, thus occluded objects are prone to being missed or misclassified.

To address these issues, we propose Spatio-Temporal Context Modeling and Refining(STCMR) for Airport Video Semantic Segmentation, a novel method featuring two dedicated modules:**Multi-scale Motion Dynamic Aggregation Module**: Models target evolution across spatio-temporal scales. It extracts multi-scale features from adjacent frames and uses a gating mechanism for the adaptive, coarse-to-fine fusion of cross-frame and cross-scale features. This enhances perception of motion information in multi-scale feature spaces, capturing Multi-scale Motion Context.**Cross-Frame Refinement Module**: uses motion-context-rich features to align adjacent frames and current-frame features to resolve occlusion. It enhances segmentation accuracy by effectively mining and fusing complementary inter-frame information.In summary, our main contributions are as follows:

1) We propose a Spatio-Temporal Context Modeling and Refining method, which can learn a unified representation of Multi-scale Motion contexts and maintain inter-frame temporal consistency.

2) A Multi-scale Motion Dynamic Aggregation Module (MMDA) is designed to explicitly model the evolution of targets across spatio-temporal scales, adaptively integrating cross-scale features through multi-level context-aware aggregation to significantly enhance motion context representation for varying object scales.

3) A Cross-Frame Refinement Module (CFR) is designed to enhance the semantic representation of the target frame features by aligning the features of adjacent frames and mining the inter-frame correlation information, while ensuring consistent segmentation results of the same target across different frames.

4) Experimental results on the AVSS dataset demonstrate the superior performance of our method over existing state-of-the-art.

II. METHODOLOGY

A. Feature Extractor

Following the feature extraction setup of Segformer [22], the feature extractor is a weight-sharing feature extractor applied to each frame in the video sequence, including the target frame V_t and reference frames $\{V_{t-l}, ..., V_{t-1}, V_t\}$, and $\{V_{t-l}, ..., V_{t-1}\}$ represent the preceding l historical frames with temporal offsets $\{l, ..., 1\}$ relative to the target frame. The feature maps are named as F=$\{F_{t-l}, ..., F_{t-1}, F_t\}$. This extractor ensures that features from different frames are aligned in the same semantic space, thereby capturing inter-frame shared semantic information(e.g., scene structures, object contours), while reducing computational redundancy.

Generally, the visual representation scale of moving objects in a scene decreases as the temporal distance to the target frame becomes shorter; conversely, it increases with greater temporal distance. Based on this spatiotemporal scale property, the features output by the Feature Extractor are reconstructed across multiple ranges along the temporal level $d \in\{\ 0,...,l\}$ using Depth-wise Convolution (DWConv), thereby generating new features Z with diverse spatiotemporal receptive fields, as shown in $Eq.(1)$:

$$Z_d = f^d\left(F_{t-d}\right) \triangleq \mathrm{GeLU}\left(\mathrm{DWConv}\left(F_{t-d}, k_{conv} = 2d+1\right)\right) \tag{1}$$

where f^d represents the inter-frame scale context encoding corresponding to the dth frame relative to the target frame, $f(\cdot)$ is the function computed by a DWConv followed by a GeLU [23] activation function, and $k_{conv} = 2d + 1$ indicates that the size of the convolution kernel is adaptively adjusted according to the temporal offset d: the larger the temporal offset, the larger the convolution kernel size. This expands the receptive field to more sufficiently capture the drastic variation information of the target scale in frames with large temporal distances.

B. Multi-scale Motion Dynamic Aggregation

To fuse these scale-varying dynamics across the temporal span and yield a more discriminative representation, we design the Motion–scale Dynamic Aggregation (MMDA) module(Fig. 3).

The inter-frame scale context encoding in $Eq.(1)$ generates l feature maps at different scale contexts. At the F_{t-d} frame relative to the target frame, the effective receptive field is $r_d = 1+\sum_{i=0}^{d}\left(k_{conv}-1\right)$ which increases as the kernel size k_{conv} grows. To capture the global context of the entire input (which is typically high-resolution), we apply global average pooling to the feature map Z_{l+1}:

$$Z_{l+1} = AvgPool(Z_l) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} Z_l(i, j, :) \tag{2}$$

where H and W denote the height and width of the feature map respectively,$Z_l(i, j, :)$ represents the channel-wise feature vector of Z_l at the pixel location.

Consequently, we jointly extract multi-scale contextual representations across $\{Z_j\}_{j=0}^{l+1}$ feature maps,$j\ \in\{\ 0,...,l+1\}$ is the number of stacking times, encapsulating both locally detailed and globally consistent cues at varying granularity levels.

To achieve dynamic merging of multi-scale feature maps, we introduce a gating selection mechanism (Selector) to dynamically control how scale features for each query are aggregated across different motion frames, thereby obtaining a unified feature representation. Specifically, we use a linear layer to compute spatial-scale aware gating weights Selector S:

$$S = \sigma\big(\mathrm{Conv}_2(\mathrm{RELU}(\mathrm{Conv}_1(F_t)))\big) \in \mathbb{R}^{H\times W\times(l+2)} \tag{3}$$

where Conv1 and Conv2 denote the first and second convolutional layers, respectively. Both adopt a 3×3 kernel size, a stride of 1, and same padding to ensure the spatial size of the output features is consistent with that of the input

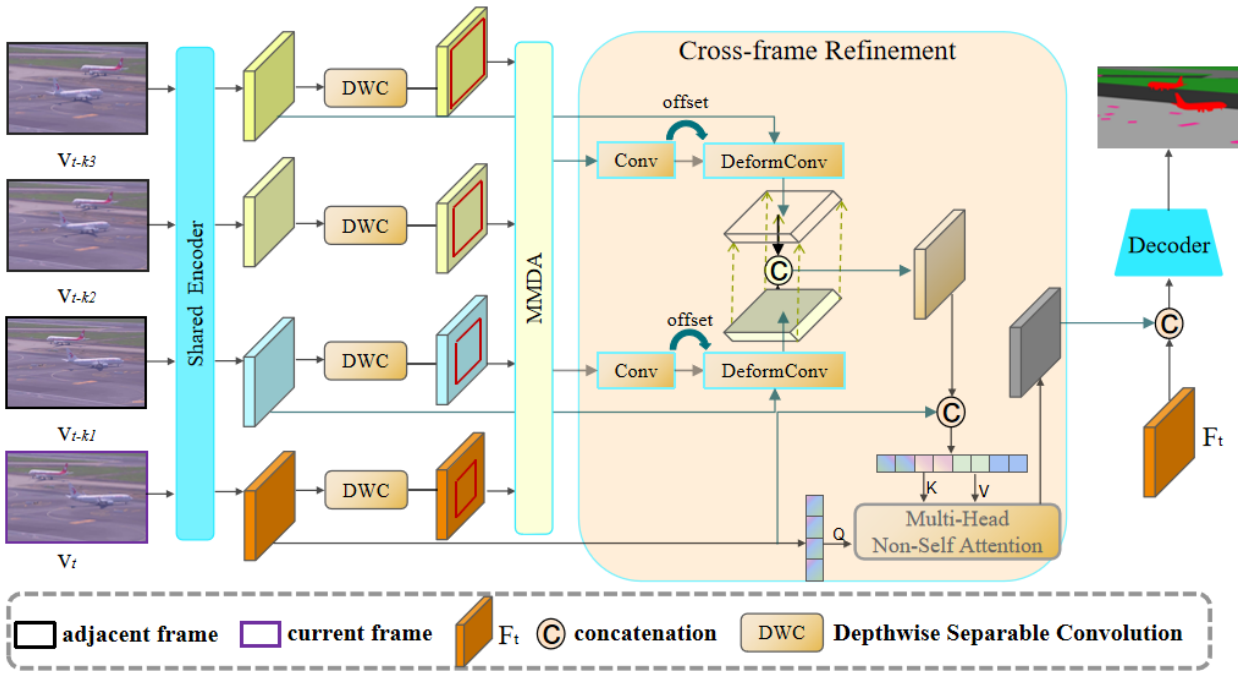

Fig. 2. Overview of the STCMR-AVSS famework. The proposed method can be divided into four modules.The Feature Extractor is a weight-sharing feature extractor applied to each frame. The Multi-scale Motion Dynamic Aggregation module(MMDA) is designed to fuse these scale-varying dynamics across the temporal span. The Cross-frame Refinement module(CFR) is designed to maintain temporal consistency by leveraging inter-frame feature correlations.Finally, The Decoder integrates the feature maps and decode the semantic segmentation output.

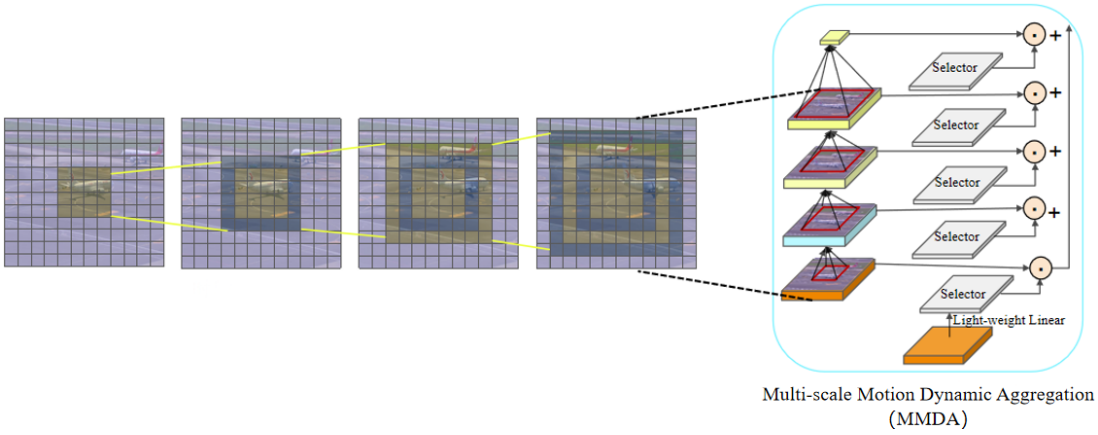

Fig. 3. Overview of the MMDA Module. As the temporal distance increases, the range of object motion becomes larger. Therefore, we expand the receptive field to more sufficiently capture the information of drastic changes in object scales in distant frames.

features. σ denotes the Sigmoid activation function, which maps the weight coefficients to the range [0,1] to quantify the contribution degree of each feature. We then perform weighted element-wise multiplication to obtain a feature map Z_{out}:

$$Z_{\text{out}} = \sum_{j=0}^{l+1} S_j \odot Z_j \in R^{H\times W\times C} \tag{4}$$

where $S_j \in R^{H\times W\times 1}$ is the jth weight matrix extracted from the gating weights S at the scale dimension. This mechanism adaptively selects the most relevant context at different scales. In this way, the different scale information from all the motion frames can be effectively aggregated.

C. Cross-frame Refinement

1) Deformable Alignment: It does not compute optical flow, but dynamically predicts the offsets of sampling convolution kernels. For the features Z_{out} generated by MMDA, we first predict a set of offsets Δp_n via a convolution layer.These offsets are positional adjustments for each sampling point of a standard convolution kernel (e.g., the 3×3 grid R).

$$\Theta = f_\theta\left(Z_{out}\right) \tag{5}$$

TABLE I
COMPARISON OF DIFFERENT METHODS USING VARIOUS MODEL

Methods	Venue	Backbone	mIoU	Weighted IoU	mAcc	mVC_8	mVC_{16}	Params (M)	FPS (f/s)
Deeplabv3 [24]	ECCV '2018	Resnet101	40.65	69.71	57.52	88.15	87.81	62.70	14.83
Pspnet [25]	CVPR '2017	Resnet101	42.34	68.54	61.12	88.36	87.15	70.50	13.29
FCN [1]	CVPR '2015	Resnet101	40.32	68.03	58.58	88.51	87.32	62.50	14.36
Dat++ [26]	CVPR '2023	Resnet101	38.75	67.89	58.90	88.28	87.11	74.00	12.65
TMANet [19]	ICIP '2021	Resnet101	39.95	69.93	55.95	87.25	85.95	76.53	10.65
SegFormer [22]	NeurIPS '2021	Mit-B1	41.04	68.34	59.49	89.06	88.12	13.80	60.34
MRCFA [20]	ECCV '2023	Mit-B1	39.95	68.20	59.11	89.82	88.61	15.40	94.34
CFFM [27]	CVPR '2023	Mit-B1	41.32	70.84	60.08	93.95	91.06	15.50	39.28
STCMR(Ours)	-	Mit-B1	**41.74**	**71.54**	**61.12**	**94.27**	**91.42**	15.45	38.75
SegFormer [22]	NeurIPS '2021	Mit-B2	41.67	68.54	61.99	89.64	88.27	24.80	55.92
MRCFA [20]	ECCV '2023	Mit-B2	41.07	68.50	59.49	89.98	88.94	16.20	95.39
CFFM [27]	CVPR '2023	Mit-B2	42.75	72.26	61.05	94.06	91.32	26.50	27.16
STCMR(Ours)	-	Mit-B2	**42.82**	**72.41**	**61.16**	**94.28**	**91.46**	26.46	27.43
SegFormer [22]	NeurIPS '2021	Mit-B5	42.71	74.13	61.99	90.96	90.84	82.10	12.35
MRCFA [20]	ECCV '2023	Mit-B5	42.81	73.72	62.73	92.54	91.34	84.50	11.06
CFFM [27]	CVPR '2023	Mit-B5	43.52	72.24	62.34	94.63	92.55	85.50	11.08
STCMR(Ours)	-	Mit-B5	**43.84**	**74.72**	**63.41**	**94.98**	**92.86**	85.48	11.11

Here, $\Theta = \{\Delta p_n \mid n = 1, \ldots, |\mathcal{R}|\}$refers to the offsets of the convolution kernels, where $\mathcal{R} = \{(-1,-1),(-1,0),(0,1),(1,1)\}$ denotes a regular grid of a 3×3 kernel. Then, the predicted offsets are adopted to perform deformable convolution on the features of support frames (i.e., the two frames closest to the target frame F_{t-min} and farthest from the target frame F_{t-max}).With Θ and Z_{out},the aligned feature $F^{'}_{t-max}$ and $F^{'}_{t-min}$

$$F^{'}_{t-max} = Deformable\,(F_{t-max}, \Theta) \tag{6}$$

$$F^{'}_{t-min} = Deformable\,(F_{t-min}, \Theta) \tag{7}$$

Instead of sampling on a regular grid, deformable convolution conducts irregular and adaptive sampling on the feature map according to the offsets Δp_n.For each position p_0 on the aligned feature map $F^{'}_{t-max}$ and $F^{'}_{t-min}$,we have:

$$F^{'}_{t-max}(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) F_{t-max}(p_0 + p_n + \Delta p_n) \tag{8}$$

$$F^{'}_{t-min}(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) F_{t-min}(p_0 + p_n + \Delta p_n) \tag{9}$$

The convolution will be operated on the irregular positions $p_n + \Delta p_n$, where the Δp_n may be fractional. To address the issue, the operation is implemented by using bilinear interpolation, which is the same as that proposed in [28].Then We concatenate $F^{'}_{t-max}$ and $F^{'}_{t-min}$ into $F^{'}_{t}$ as follows:

$$Z'_{out} = concat\big(F'_{t-max}, F'_{t-min}\big) \tag{10}$$

It aligns all object features in our adjacent frames.

2) Multi-head Non-self Attention: In order to mine useful information from neighboring frames,we use a nonself attention mechanism.Unlike the traditional self-attention mechanism that computes the query, key, and value from the same input, our non-self attention mechanism utilizes different inputs to calculate the query, key, and value.For the i-th window partition in F_t , the query Q_i, key K_i, and value V_i are computed using three fully connected layers as follows:

$$Q_i = \mathrm{FC}(F_t[i]), \quad K_i = \mathrm{FC}(Z'_{out}), \quad V_i = \mathrm{FC}(Z'_{out}) \tag{11}$$

where $FC(\cdot)$ represents a FC layer. Next, we use non-self attention to update the target frame features, given by:

$$\mathbf{F}'_t[i] = \mathrm{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{c}} + B\right) V_i \tag{12}$$

where B represents the position bias, following [29]. Note that we omit the formulation of the multi-head attention [30] for simplicity.

D. Lightweight Decoder

The lightweight decoder consists of four main steps. First, features *F* = concat(F'_t,F_t) go through an MLP layer to unify the channel dimension. Then, in a second step, features are upsampled to $\frac{1}{4}$th and concatenated together, Thirdly, a MLP layer is adopted to fuse the concatenated features Z_t. Finally, another MLP layer takes the fused feature to predict the segmentation mask *M* with a $\frac{H}{4} \times \frac{W}{4} \times N_{cls}$ resolution, where N_{cls} is the number of categories.

III. EXPERIMENTS

A. Datasets, Evaluation and Experimental Details

To the best of our knowledge, the AVSS dataset is currently the only video semantic segmentation dataset dedicated to airport scenes. It comprises finely annotated surveillance video clips covering various airport environments and identifies 18 major semantic categories, serving as the primary benchmark for airport video semantic segmentation. To evaluate segmentation performance on this dataset, we employ Mean Intersection-over-Union (mIoU) and Mean Class Accuracy (mAcc) as standard metrics. In addition, drawing on methodologies from prior work, we introduced a weighted IoU to further assess segmentation quality and proposed a Video Consistency (VC) metric to evaluate model stability across

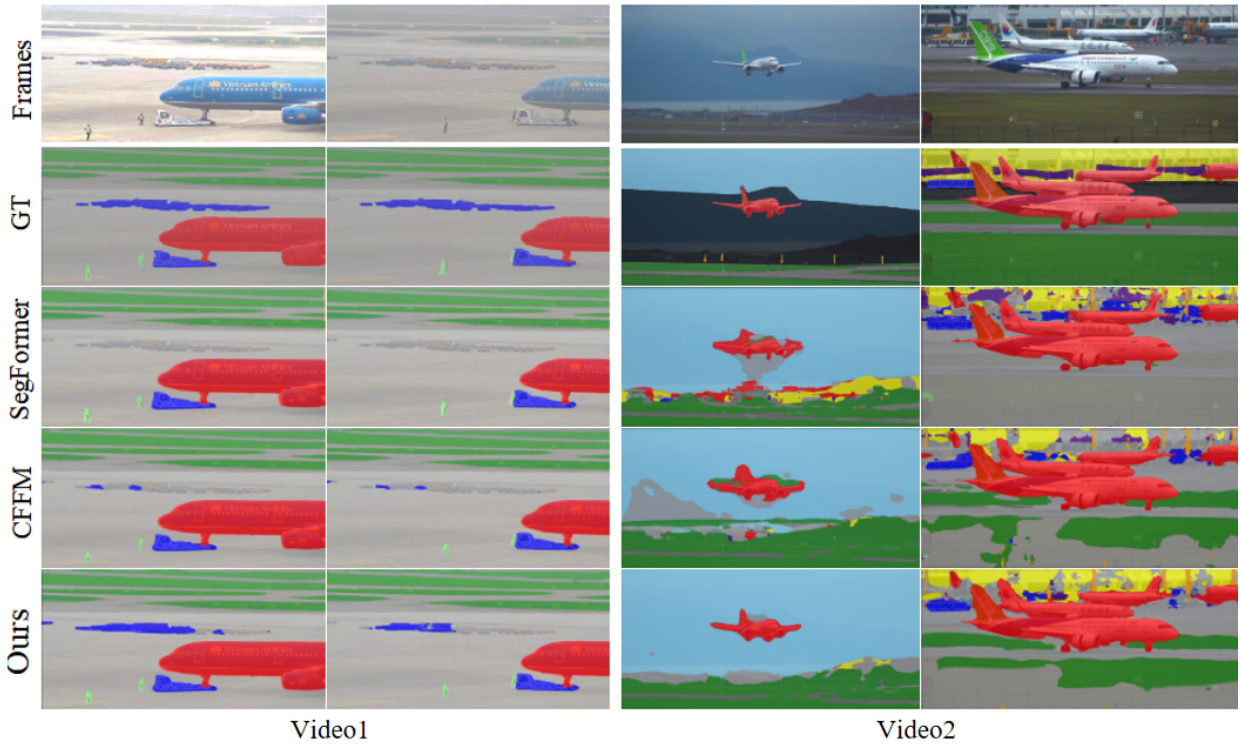

Fig. 4. Segmentation Visualization in Airport Scenes with Substantial Severe Occlusion and Scale Variations. Qualitative results show that the proposed method preserves boundaries more completely and reduces misjudgments in occluded scenarios (e.g., cars in Video 1), and maintains robust segmentation under significant scale changes (e.g., aircraft landing in Video 2).

consecutive video frames. We conducted experiments using an RTX 4090 GPU for 16000 iters using the AdamW optimizer.We use the crop size of 480 × 480 for the AVSS. The batch size is set to 4 with an initial learning rate of 1×10^{-6}.The CrossEntropyLoss weights are set 1. Unless otherwise specified, our model utilizes three reference frames.

B. Comparison with State-of-the-art Methods

To verify the effectiveness of STCMR-AVSS, we conduct comparative experiments on the AVSS dataset against two categories of methods: static image semantic segmentation methods [1], [22], [24]–[26], and video semantic segmentation methods [19], [20], [27].Evaluation metrics include core segmentation metrics (mIoU, Weighted IoU, mAcc) and video consistency metrics (mVC8, mVC16), along with model efficiency indicators (Params, FPS). All results are summarized in Table I. For ResNet101-backbone methods, PSPNet achieves the highest mIoU 42.34% but with a low Weighted IoU 68.54% and FPS (13.29 f/s). TMANet, a video segmentation method, only reaches 39.95% mIoU and 10.65 f/s. In contrast, our STCMR-AVSS with the Mit-B1 backbone achieves 41.74% mIoU and 71.54% Weighted IoU, outperforming all ResNet101-based methods in segmentation quality while maintaining higher inference efficiency.

We further compare STCMR-AVSS with advanced video segmentation methods using Mit-series backbones,for Mit-B1: STCMR-AVSS achieves 41.74% mIoU, 0.7% higher than SegFormer 41.04%, with mVC8/mVC16 reaching 94.27%/91.42%—surpassing SegFormer by 3.21%/3.30%. Its Params (15.45M) and FPS (38.75 f/s) are comparable to MRCFA, balancing performance and efficiency;for Mit-B5:STCMR-AVSS achieves the best overall performance with 43.84% mIoU, 74.72% Weighted IoU, and 94.98% mVC8, outperforming CFFM (43.52% mIoU, 94.63% mVC8) by 0.32% and 0.35% respectively. Across all backbones, STCMR-AVSS consistently outperforms the corresponding baselines, demonstrating the stability and effectiveness of the proposed modules. This improvement substantiates the importance of leveraging temporal dependencies and multi-scale feature correlations to enhance semantic segmentation accuracy in airport surveillance videos.

TABLE II
ABLATION STUDY OF COMPONENTS WITH DIFFERENT BACKBONES.

Methods			backbone	mIoU	mVC_{16}	mAcc
Base	MMDA	CFR				
√	×	×	Mit-B1	41.04	88.12	59.49
√	√	×	Mit-B1	41.93	90.93	61.04
√	×	√	Mit-B1	41.15	89.95	60.02
√	√	√	Mit-B1	**41.74**	**91.42**	**61.12**
√	×	×	Mit-B5	42.71	90.96	61.99
√	√	×	Mit-B5	43.71	91.98	62.26
√	×	√	Mit-B5	42.93	91.29	60.14
√	√	√	Mit-B5	**43.84**	**92.86**	**63.41**

C. Ablation Study and Visualization

To validate the effectiveness of core components in the STCMR framework and demonstrate its advantages in airport video semantic segmentation, we design four experimental settings on MiT-B1 and MiT-B5 backbones: Base (SegFormer-based baseline), Base+MMDA, Base+CFR, and STCMR (our model with both modules). All experiments adopt consistent parameters for fairness, and results are summarized in Table II.The full STCMR-AVSS model achieved the highest segmentation accuracy with a 1.13% mIoU gain, demonstrating the complementary and synergistic effects of both modules. Then,Visualization (Fig. 4) on challenging samples (occlusion/scale variation) confirms STCMR's advantages.Compared with SOTA methods (e.g., CFFM), STCMR better handles occlusion (accurate segmentation of occluded aircraft/vehicles), adapts to aircraft scale changes (complete contour retention), and maintains inter-frame semantic consistency (minimal segmentation jitter).

IV. Conclusion

This paper presents the STCMR-AVSS framework, a spatio-temporal context modeling and refining approach for airport video semantic segmentation, where its core MMDA module uses a gating mechanism to adaptively fuse multi-scale motion context to address drastic scale variations in airport scenes and its CFR module explores temporal consistency from adjacent frames to improve accuracy in heavily occluded regions, with the two modules cooperating synergistically to overcome traditional method limitations, extensive experiments on the AVSS benchmark confirming its superior performance over state-of-the-art methods across multiple metrics, and future work focusing on enhancing generalization to other scenes and optimizing inference speed for real-time applications.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[3] X. Li, F. Xu, A. Yu, X. Lyu, H. Gao, and J. Zhou, "A frequency decoupling network for semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[4] W. M. Elmessery, D. V. Maklakov, T. M. El-Messery, D. A. Baranenko, J. Gutiérrez, M. Y. Shams, T. A. El-Hafeez, S. Elsayed, S. K. Alhag, F. S. Moghanm *et al.*, "Semantic segmentation of microbial alterations based on segformer," *Frontiers in Plant Science*, vol. 15, p. 1352935, 2024.

[5] Z. Ni, X. Chen, Y. Zhai, Y. Tang, and Y. Wang, "Context-guided spatial feature reconstruction for efficient semantic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 239–255.

[6] Y. Wang, W. Zhao, C. Cao, T. Deng, J. Wang, and W. Chen, "Sfpnet: Sparse focal point network for semantic segmentation on general lidar point clouds," in *European Conference on Computer Vision*. Springer, 2024, pp. 403–421.

[7] Q. Cao, Y. Chen, C. Ma, and X. Yang, "Few-shot rotation-invariant aerial image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2023.

[8] P. Lei and S. Todorovic, "Recurrent temporal deep field for semantic video labeling," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 302–317.

[9] D. Nilsson and C. Sminchisescu, "Semantic video segmentation by gated recurrent flow propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6819–6828.

[10] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi, "Temporally distributed networks for fast video semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8815–8824.

[11] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 408–417.

[12] R. Gadde, V. Jampani, and P. V. Gehler, "Semantic video cnns through representation warping," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4453–4462.

[13] S. Liu, C. Wang, R. Qian, H. Yu, R. Bao, and Y. Sun, "Surveillance video parsing with single frame supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 413–421.

[14] Y. Liu, C. Shen, C. Yu, and J. Wang, "Efficient semantic video segmentation with per-frame inference," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 352–368.

[15] Y. Weng, M. Han, H. He, M. Li, L. Yao, X. Chang, and B. Zhuang, "Mask propagation for efficient video semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 7170–7183, 2023.

[16] Y. Zheng, H. Yang, and D. Huang, "Deep common feature mining for efficient video semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[17] B. Li, T. Yan, W. He, and X. Zhu, "Stanet: a spatial-temporal aggregation network for video deraining," in *2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE)*. IEEE, 2023, pp. 470–479.

[18] Y. Li, W. Wang, W. Chen, L. Niu, J. Si, C. Qian, and L. Zhang, "Video semantic segmentation via sparse temporal transformer," in *Proc. ACM Int. Conf. Multimed.*, 2021, pp. 59–68.

[19] H. Wang, W. Wang, and J. Liu, "Temporal memory attention for video semantic segmentation," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 2254–2258.

[20] G. Sun, Y. Liu, H. Tang, A. Chhatkuli, L. Zhang, and L. van Gool, "Mining relations among cross-frame affinities for video semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2023, pp. 522–539.

[21] G. Sun, Y. Liu, H. Ding, M. Wu, and L. Van Gool, "Learning local and global temporal contexts for video semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 10, pp. 6919–6934, 2024.

[22] 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.

[23] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, and P. H. S. Torr, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[24] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[26] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Dat++: Spatially dynamic vision transformer with deformable attention," *arXiv preprint arXiv:2309.01430*, 2023.

[27] G. Sun, Y. Liu, H. Ding, T. Probst, and L. van Gool, "Coarse-to-fine feature mining for video semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3126–3137.

[28] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[30] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.