GasTwinFormer: A Hybrid Vision Transformer for Livestock Methane Emission Segmentation and Dietary Classification in Optical Gas Imaging

Toqi Tahamid Sarker[1], Mohamed Embaby[2], Taminul Islam[1], Amer AbuGhazaleh[1], Khaled Ahmed[1]
[1]Southern Illinois University Carbondale, USA [2]University of California, Davis, USA
{toqitahamid.sarker, taminul.islam, aabugha, khaled.ahmed}@siu.edu, membaby@ucdavis.edu

Abstract

Livestock methane emissions represent 32% of human-caused methane production, making automated monitoring critical for climate mitigation strategies. We introduce GasTwinFormer, a hybrid vision transformer for real-time methane emission segmentation and dietary classification in optical gas imaging through a novel Mix Twin encoder alternating between spatially-reduced global attention and locally-grouped attention mechanisms. Our architecture incorporates a lightweight LR-ASPP decoder for multi-scale feature aggregation and enables simultaneous methane segmentation and dietary classification in a unified framework. We contribute the first comprehensive beef cattle methane emission dataset using OGI, containing 11,694 annotated frames across three dietary treatments. GasTwinFormer achieves 74.47% mIoU and 83.63% mF1 for segmentation while maintaining exceptional efficiency with only 3.348M parameters, 3.428G FLOPs, and 114.9 FPS inference speed. Additionally, our method achieves perfect dietary classification accuracy (100%), demonstrating the effectiveness of leveraging diet-emission correlations. Extensive ablation studies validate each architectural component, establishing GasTwinFormer as a practical solution for real-time livestock emission monitoring. Please see our project page at gastwinformer.github.io.

1. Introduction

Methane (CH_4) represents a potent greenhouse gas with a global warming potential 84 times greater than carbon dioxide over a 20-year timeframe [14]. Agriculture accounts for 40% of human-caused methane emissions, with livestock responsible for roughly 32% [21]. As global food demand is expected to increase by 70% by 2050, developing efficient monitoring systems for livestock methane emissions has become critical for climate mitigation [20].

The relationship between livestock diet composition and methane production creates opportunities for integrated monitoring systems. Different feed regimens significantly influence emission patterns—high-forage diets typically increase methane production due to fiber fermentation, while grain-rich diets can reduce emissions [4]. Advanced monitoring technologies enable precise quantification of these emission patterns, creating opportunities for data-driven livestock management through real-time assessment of feeding strategies and emission mitigation interventions [13].

Traditional methane quantification methods rely on respiration chambers or emission factor calculations, which suffer from high costs, labor-intensive protocols, and inability to capture real-time dynamics [22]. Recent advances in optical gas imaging (OGI) offer non-invasive, continuous monitoring capabilities using thermal infrared cameras operating in the 7-8.5 μm spectral range [25, 27]. However, OGI presents computational challenges including low signal-to-noise ratios, complex thermal backgrounds, and irregular plume morphology requiring automated analysis [7].

Vision transformers have revolutionized dense prediction tasks through global context modeling, but face computational challenges with high-resolution OGI data due to quadratic attention complexity [9]. Recent hybrid attention mechanisms show promise for balancing efficiency with representational capacity, but have not been adapted for gas plume segmentation [6, 8, 31].

We propose a novel architecture called GasTwinFormer for semantic segmentation of methane emissions and dietary treatment classification in beef cattle OGI camera images. In the encoding stage, we develop a Mix Twin encoder that combines efficient multi-head attention (EMA) [30] with locally-grouped self-attention (LSA) [2] to capture both global context and local details for precise gas plume detection. This dual attention approach enables effective processing of thermal infrared imagery while maintaining computational efficiency. In the decoding stage, we use a hierarchical LR-ASPP decoder [5] that processes features from multiple encoder stages to generate accurate segmentation predictions. Our framework performs both pixel-wise methane segmentation and dietary classification using

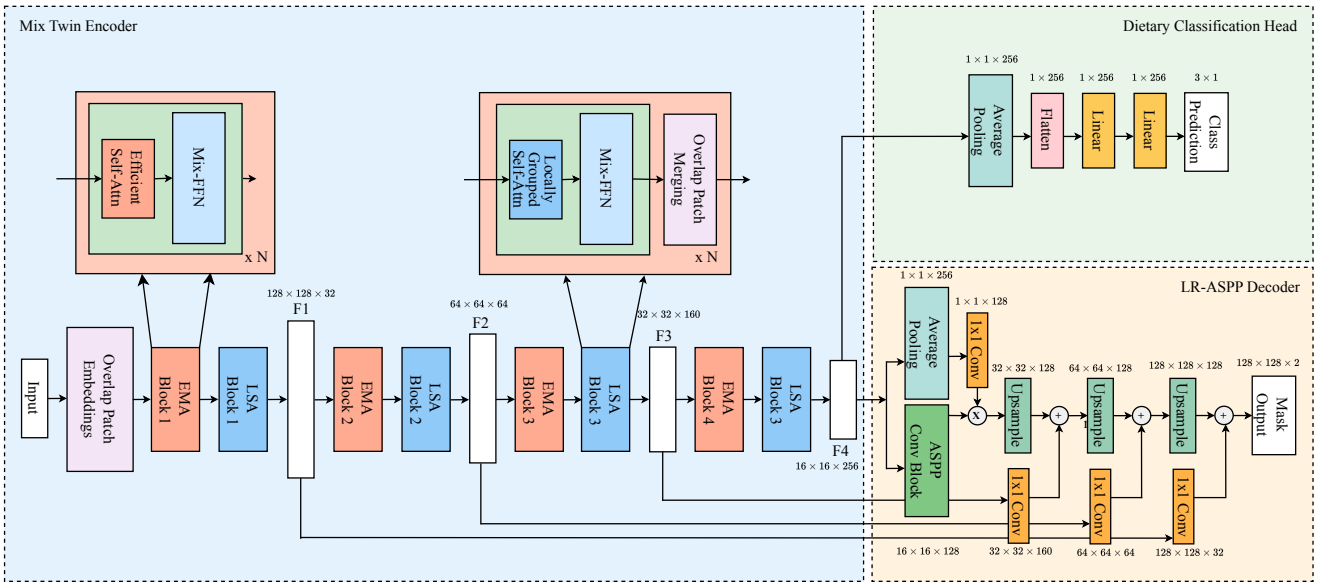

Figure 1. GasTwinFormer architecture. The Mix Twin encoder uses alternating EMA and LSA blocks across four hierarchical stages. The LR-ASPP decoder performs methane segmentation while a separate classification head predicts dietary treatment.

shared features. The main contributions of this study are as follows:

1. GasTwinFormer, a hybrid transformer-based architecture that enables concurrent methane plume segmentation and dietary treatment classification in livestock monitoring applications.
2. A comprehensive beef cattle methane emission dataset captured through OGI technology, comprising 11,694 semi-automatically annotated frames spanning three feeding regimens.
3. Extensive benchmarking and performance analysis against existing state-of-the-art segmentation approaches, demonstrating superior accuracy and computational efficiency across multiple metrics.

2. Related Work

Optical Gas Imaging and Methane Detection. Wang *et al.* [25] pioneered computer vision for methane detection using infrared cameras, developing GasNet with 95% detection accuracy on ∼1M labeled frames. VideoGasNet [26] extended this work to leak size classification using 3D CNNs. Recent advances include vision transformers for satellite methane detection [18] and CNNs for airborne emission quantification [7]. Most recently, Sarker *et al.* [19] introduced Gasformer, achieving 88.56% mIoU on livestock datasets using Mix Vision Transformer encoders. However, existing approaches lack systematic integration of global and local attention for enhanced boundary delineation in challenging thermal imagery.

Vision Transformers for Dense Prediction. Dosovitskiy *et al.* [3] established Vision Transformers for image classification, while Ranftl *et al.* [17] introduced Dense Vision Transformers for dense prediction tasks. Hierarchical designs have proven effective: Swin Transformer [10] uses shifted windowing for computational efficiency, PVT [28] establishes hierarchical principles through progressive spatial reduction, and SegFormer [30] achieves state-of-the-art performance (51.8% mIoU on ADE20K) through efficient MLP decoders and Mix Vision Transformers with spatial inductive bias.

Hybrid Attention Mechanisms. Chu *et al.* [2] proposed Twins architectures (PCPVT and SVT) that systematically combine different attention mechanisms. Twins introduces Locally-Grouped Self-Attention (LSA) partitioning spatial dimensions into non-overlapping windows for linear complexity, while maintaining local pattern recognition. Yang *et al.* [31] demonstrated that treating global and local attention as complementary achieves superior dense prediction performance. However, existing hybrid approaches focus on natural images and have not been adapted for gas plume segmentation challenges.

Research Gaps. Despite advances in methane detection and vision transformers, three critical limitations hinder practical OGI-based livestock monitoring. Current limitations in OGI-based livestock monitoring include: **(1)** reliance on single-scale attention mechanisms that inadequately balance global context and local precision for gas plume characteristics, **(2)** treatment of methane detection as an isolated task without leveraging established diet-emission correlations, and **(3)** absence of comprehensive livestock-specific datasets capturing real-world farming complexities beyond controlled laboratory conditions. Our GasTwinFormer addresses these limitations through

hybrid attention design, multi-task learning integration, and comprehensive dataset development.

3. Method

GasTwinFormer consists of three primary components: (1) a hierarchical Mix Twin encoder that combines efficient multi-head self-attention (EMA) from SegFormer's Mix Transformer [30] with locally-grouped self-attention (LSA) from Twins [2], (2) a hierarchical lightweight reduced Atrous Spatial Pyramid Pooling decoder (LR-ASPP) [5] for multi-scale feature aggregation and pixel-wise methane segmentation, and (3) a dietary classification head for scene-level prediction. Figure 1 illustrates the complete architecture pipeline.

3.1. Mix Twin Encoder

Hybrid Attention Architecture. The backbone encoder follows a hierarchical design with four stages that progressively reduce spatial resolution from $H/4$ to $H/32$ while expanding channel capacity from 32 to 256. Within each stage, we use an EMA→LSA composition: an EMA block establishes global relationships via spatially reduced attention, followed by an LSA block that refines local structure using 5×5 windows. Each stage therefore contains exactly one EMA–LSA pair (EL). SegFormer uses only efficient (EMA) attention in every block. Twins–PCPVT uses only global sub-sampled attention (GSA). Twins–SVT places LSA first and GSA second in a repeating LSA→GSA sequence. Accordingly, composing EL within each stage captures long-range plume context via EMA and sharp boundary details via LSA in a single pass, yielding higher accuracy than EMA-only (SegFormer), GSA-only (Twins–PCPVT), or LSA→GSA alternation (Twins–SVT); quantitative gains are reported in Tab. 2.

Hierarchical Multi-Scale Feature Extraction. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the encoder generates multi-scale feature representations $\{F_1, F_2, F_3, F_4\}$ with progressive spatial downsampling and corresponding channel dimensions $\{32, 64, 160, 256\}$, respectively. The EMA-LSA pairs in each stage incorporate overlapped patch embedding and layer normalization, resulting in 8 total blocks across the four-stage encoder.

Overlapped Patch Embedding. We use overlapped patch embedding to preserve spatial continuity for precise boundary localization. The first stage uses a 7×7 convolution with stride 4 and padding 3, while subsequent stages use 3×3 convolutions with stride 2 and padding 1 for efficient downsampling.

Efficient Multi-Head Attention. Standard multi-head self-attention mechanisms exhibit quadratic computational complexity $O(N^2)$ with respect to spatial resolution N = H × W, creating computational bottlenecks for high-resolution dense prediction tasks. We address this limitation by adopting the Efficient Multi-Head Attention (EMA) from SegFormer [30], which builds upon the spatial reduction process introduced in Pyramid Vision Transformer [30]. This approach reduces complexity to $O(N^2/R)$ through spatial reduction of key and value representations while maintaining full-resolution queries. For each stage i with reduction ratio R_i, both key and value matrices are spatially downsampled to dimensions $\mathbb{R}^{(N/R_i)\times C}$ using convolutions with kernel size and stride equal to R_i. The attention computation becomes:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}', \mathbf{V}') = \text{Softmax}\left(\frac{\mathbf{Q}(\mathbf{K}')^T}{\sqrt{d_{\text{head}}}}\right)\mathbf{V}' \tag{1}$$

where $\mathbf{K}'$ and $\mathbf{V}'$ are the spatially reduced key and value representations with dimensions $\mathbb{R}^{(N/R_i)\times C}$.

We use stage-adaptive reduction ratios $R = \{8, 4, 2, 1\}$ that align with the hierarchical nature of feature learning. Early stages utilize aggressive reduction ($R = 8$) to handle high-resolution features efficiently, while later stages progressively decrease reduction ratios as spatial dimensions naturally diminish through downsampling. This strategy ensures computational tractability in high-resolution stages while maintaining fine-grained attention capabilities in semantically rich later stages.

Locally-Grouped Self-Attention. While EMA achieves computational efficiency through spatial reduction, it may compromise fine-grained spatial detail preservation that is critical for accurate boundary delineation in methane plume segmentation. To address this limitation, we integrate LSA from Twins-SVT [2] as the second component in our hybrid attention pattern. LSA complements the global context modeling of efficient attention by capturing fine-grained local structures through spatially partitioned attention computation.

The LSA addresses the quadratic complexity challenge through spatial partitioning rather than spatial reduction. Given an input feature map $\mathbf{X} \in \mathbb{R}^{B\times N\times C}$ where $N = H \times W$, LSA partitions the spatial dimensions into non-overlapping windows of size $w_1 \times w_2$. Self-attention is then computed independently within each local window:

$$\text{LSA}(\mathbf{X}) = \text{Concat}_{i,j}\left(\text{Attention}(\mathbf{X}_{i,j})\right) \tag{2}$$

where $\mathbf{X}_{i,j} \in \mathbb{R}^{B\times w_1 w_2\times C}$ represents the feature tokens within window (i, j), and the concatenation operates over all $\lceil H/w_1 \rceil \times \lceil W/w_2 \rceil$ windows. The attention computation within each window follows the standard formulation:

$$\text{Attention}(\mathbf{X}_{i,j}) = \text{Softmax}\left(\frac{\mathbf{Q}_{i,j}\mathbf{K}_{i,j}^T}{\sqrt{d_{\text{head}}}}\right)\mathbf{V}_{i,j} \tag{3}$$

This design achieves computational complexity of $\mathcal{O}(w_1 w_2 HWd)$, which scales linearly with spatial resolution since the window size $w_1 w_2$ remains fixed. For our implementation with $w_1 = w_2 = 5$, the complexity becomes $\mathcal{O}(25HWd)$, providing substantial efficiency gains

while maintaining sufficient receptive field coverage for local pattern recognition.

Mix Feed-Forward Network. Both Transformer Block and LSA Block utilize the Mix Feed-Forward Network (Mix-FFN) module from SegFormer [30], which eliminates the need for explicit positional encodings while providing spatial inductive bias. Unlike Vision Transformers that use fixed-resolution positional encodings, we argue that positional encoding is not necessary for dense prediction tasks. Instead, Mix-FFN considers the effect of zero padding to leak location information by directly incorporating a 3×3 convolution in the feed-forward network. The Mix-FFN operation is formulated as:

$$\text{Mix-FFN}(\mathbf{x}) = \text{MLP}(\text{GELU}(\text{Conv}_{3\times3}(\text{MLP}(\mathbf{x})))) + \mathbf{x} \quad (4)$$

where $\mathbf{x}$ is the feature from the self-attention module. Mix-FFN mixes a 3×3 convolution and MLPs into each feed-forward network. The 3×3 convolution is sufficient to provide positional information for transformers through the spatial connectivity and zero-padding effects. We use depth-wise convolutions to reduce the number of parameters and improve computational efficiency.

3.2. Hierarchical LR-ASPP Decoder

For dense prediction tasks, we use a lightweight decoder that efficiently aggregates multi-scale features from our hierarchical encoder. Building upon LR-ASPP from MobileNetV3 [5], we propose an adaptive variant that accommodates variable input resolutions while maintaining computational efficiency. Our Hierarchical LR-ASPP decoder processes multi-scale features $\{F_1, F_2, F_3, F_4\}$ through two parallel pathways: F_4 features are processed through the main ASPP path, while F_1, F_2, F_3 features are processed through dedicated 1×1 convolution branches. The operations are:

$$\begin{aligned} F_{\text{pool}} &= \text{Sigmoid}(\text{Conv}_{1\times1}(\text{AdaptiveAvgPool}(F_4))) \\ F_{\text{aspp}} &= \text{Conv}_{1\times1}(F_4) \odot \text{Upsample}(F_{\text{pool}}) \\ F_{\text{branch}_i} &= \text{Conv}_{1\times1}(F_i), \quad i \in \{1,2,3\} \\ F_{\text{out}} &= \text{ProgressiveFusion}(F_{\text{aspp}}, \{F_{\text{branch}_3}, F_{\text{branch}_2}, F_{\text{branch}_1}\}) \end{aligned} \quad (5)$$

where $\odot$ denotes element-wise multiplication, and progressive fusion sequentially upsamples, concatenates, and fuses features from deeper to shallower levels. This design preserves both semantic information from deep features and spatial details from shallow features essential for accurate methane plume boundary delineation.

3.3. Dietary Classification Head

To enable simultaneous scene-level classification alongside dense plume segmentation, we incorporate a lightweight classification head that processes the highest-level semantic features from the encoder. The classification head employs a simple yet effective architecture consisting of adaptive average pooling, followed by a two-layer fully connected network with ReLU activation and dropout regularization. The classification head operates on the final stage features F_4 to predict dietary treatment categories: High Forage (HF), Mixed Diet (MD), and High Grain (HG).

3.4. Gaussian Plume Weighted Dice Loss

We incorporate the Gaussian Plume Weighted Dice Loss [36] to leverage physical constraints from gas dispersion behavior in our segmentation framework. This loss function addresses the inherent characteristics of gas plume dynamics by applying spatially-varying weights based on the Gaussian plume model. The loss formulation applies pixel-wise weights according to the Gaussian distribution:

$$w(p) = \exp\left(-\frac{(p_x - \mu_x)^2}{2\sigma_x^2} - \frac{(p_y - \mu_y)^2}{2\sigma_y^2}\right) \quad (6)$$

where $p = (p_x, p_y)$ denotes pixel coordinates, (μ_x, μ_y) represents the plume center computed via center-of-mass on predicted masks, and (σ_x, σ_y) denote the horizontal and vertical diffusion scales estimated through weighted standard deviation with adaptive bounds $[\frac{W}{20}, \frac{W}{2}]$ and $[\frac{H}{20}, \frac{H}{2}]$ respectively, where W and H are the image dimensions. The weighted Dice loss is then computed as:

$$L_{\text{weighted}} = 1 - \frac{2\sum_p w(p) \cdot y_p \cdot \hat{y}_p + \epsilon}{\sum_p w(p) \cdot y_p + \sum_p w(p) \cdot \hat{y}_p + \epsilon} \quad (7)$$

where y_p and $\hat{y}_p$ represent the ground truth and predicted segmentation values at pixel p, and ϵ is a small constant for numerical stability.

4. Beef Cattle Methane Emission Dataset

We present a comprehensive dataset for methane emission detection from beef cattle, captured using this OGI camera. This dataset addresses the critical need for computer vision benchmarks in livestock emission monitoring, particularly for developing and evaluating segmentation algorithms under challenging real-world conditions. We use the FLIR Gx320 OGI camera for methane emission detection. The camera operates in the 3.2–3.4 μm spectral range, optimized for hydrocarbon detection through mid-wave infrared sensing. Key specifications include 320×240 pixel resolution and <10 mK thermal sensitivity. The camera detects methane concentrations as low as 9.6 ppm·m under optimal conditions with 10°C thermal contrast [23].

In Vivo Trial Design. The primary aim of this in vivo trial was to assess the efficacy of combining optical gas imaging with deep learning to detect and segment methane emissions from ruminant animals across different dietary treatments. The study utilized twelve postpartum beef cows (1200 lb ± 23) over a 30-day period, with 4 animals assigned to one of three dietary treatment groups. Each group was housed and fed together in separate feed stalls, receiving 30 lb of diet mix per cow daily. All cows received

Diet Type	Images	Percentage	Videos	Train (70%)	Val (15%)	Test (15%)
High Forage	2,730	23.4%	10	1,906	404	420
Mixed Diet	4,658	39.8%	5	3,258	696	704
High Grain	4,306	36.8%	4	3,013	644	649
Total	**11,694**	**100.0%**	**19**	**8,177**	**1,744**	**1,773**

Table 1. Beef cattle methane emission dataset statistics. Percentages show distribution within 11,694 annotated frames.

feed twice daily at 7 AM and 7 PM, where hay was offered first, followed by the grain mix. All animals had free access to clean water and were housed at Southern Illinois University's beef center barns, managed in accordance with Institutional Animal Care and Use Committee guidelines (protocol number 22-016) [12]. We conducted controlled experiments across three dietary treatment groups to investigate the relationship between feed composition and methane emissions: High Forage Group (HF) fed 100% hay consisting of grass and legume mix; Mixed Diet Group (MD) fed 50% hay mix and 50% grain mix (67.5% corn, 25% DDGS, and 7.5% mineral mix); and High Grain Group (HG) fed 20% hay mix and 80% grain mix, with grain levels increased gradually to prevent acidosis and facilitate adaptation to the high-grain diet. Every cow was kept in an animal chute for 20 minutes for gas recording, and at the end of the experiment, all cows were moved to the holding barn two hours after morning feeding. The gas imaging was performed using the TELEDYNE FLIR Gx320, with the infrared camera mounted in a lateral position approximately 4 feet from the cow's head. Following recording, cows were returned to their assigned barn. The same recording procedure was repeated the next day to collect additional data required for model training.

Image Acquisition. Videos were captured in black-hot thermal mode (dark gas plumes against light backgrounds) in FLIR's CSQ format with 14-bit radiometric data. We converted CSQ files to MP4 using FLIR Thermal Studio, then extracted frames as 8-bit grayscale PNG images (0-255 intensity values with three identical channels).

Dataset Statistics and Composition. Our dataset comprises 208,149 frames extracted at 30 fps from 19 FLIR thermal recordings across dietary treatment groups. Each frame has 640×480 pixel resolution and is stored as 8-bit grayscale PNG files with values ranging 0-255. We identified and annotated 11,694 frames (5.6% of 208,149 frames) containing visible methane plumes, reflecting the intermittent nature of bovine eructation events. As shown in Tab. 1, the annotated frames are distributed across dietary treatments as: 4,658 mixed diet (39.8%), 4,306 high grain (36.8%), and 2,730 high forage (23.4%) frames. This distribution reflects both biological emission differences and collection constraints across treatments. For model development, we employed temporal splitting to preserve emission sequence integrity: 70% of consecutive frames for training, 15% for validation, and 15% for testing within each video. Table 1 details the resulting splits: 8,177 training frames, 1,744 validation frames, and 1,773 test frames. This ensures evaluation on future time points relative to training data, providing realistic generalization assessment. All 19 videos contribute to each split while maintaining dietary treatment proportions. We excluded the remaining ∼196k non-plume frames to avoid severe class imbalance without meaningful segmentation training signal.

Annotation Methodology. We developed a multi-stage annotation pipeline combining classical image processing, deep learning, and manual refinement to generate reliable ground truth masks for ephemeral methane plumes with low contrast and irregular morphology. Our pipeline consists of three complementary approaches: (1) *Classical processing* employs temporal background subtraction using exponential moving average over 5 frames, followed by motion masking (thresholds 20-60), adaptive mean thresholding (block sizes 300-5001, constants 5-15), watershed segmentation with Sobel edge detection, and morphological refinement with size filtering (>2000 pixels) and eccentricity filtering (>0.95) to remove linear artifacts. (2) *Deep learning processing* uses a Gasformer [19] model trained on initial classical masks to identify subtle patterns beyond traditional methods. (3) *Enhanced processing* applies CLAHE, intensity rescaling, and non-local means denoising ($h = 15$) for improved candidate generation. For each frame, we generate three mask candidates from these approaches and perform manual inspection to select the most accurate representation using contrast-enhanced overlays. Figure 2 (second column) shows the resulting ground truth masks for all three dietary treatments.

5. Results

5.1. Implementation Details

We implement all experiments using PyTorch and MMSegmentation framework on a server with Intel Xeon Gold 6338 (2.00GHz), NVIDIA A100 80GB GPU, and 512GB RAM. We evaluate GasTwinFormer against comprehensive baselines spanning transformer-based architectures (SegFormer [30], Twins PCPVT-S [2], Twins SVT-S [2], Gasformer [19], iFormer [35]) and CNN-based methods (DeepLabV3 [1], BiSeNetV1 [32], Fast-FCNN [16], ICNet [34], UperNet [29], BiSeNetV2 [33], DDRNet [15], RepViT [24]), with all models utilizing ImageNet pretrained weights where available. For the main results, GasTwinFormer uses the EL-EL-EL-EL hybrid attention pattern with 5×5 LSA window size and Gaussian Plume Weighted Dice Loss, as determined optimal through ablation studies in Sec. 5.3. Training proceeds for 80,000 iterations using AdamW optimizer (learning rate 6×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01) with linear warmup from 10^{-6} over 1,500 iterations followed by

Method	Backbone	mIoU (%)↑	mF1 (%)↑	Diet Acc (%)↑	Diet F1 (%)↑	Params (M)↓	FLOPs (G)↓	FPS ↑	Year
Transformer-based Methods									
SegFormer [30]	MiT-B0	72.11	81.57	100.0	100.0	3.782	7.885	119.66	2021
Twins [2]	PCPVT-S	74.05	83.25	100.0	100.0	27.906	44.34	61.60	2021
Twins [2]	SVT-S	72.06	81.62	100.0	100.0	27.846	38.471	51.64	2021
GasFormer [19]	MiT-B0	72.25	81.69	100.0	100.0	3.716	9.913	102.29	2024
iFormer [35]	iFormer-T	65.99	75.87	99.77	99.80	6.804	24.267	113.83	2025
CNN-based Methods									
DeepLabV3 [1]	ResNet-50	70.36	80.03	100.0	100.0	68.625	270.0	91.79	2017
BiSeNetV1 [32]	ResNet-18	52.87	59.29	95.32	95.76	13.455	14.821	**243.07**	2018
Fast-FCNN [16]	FastSCNN	54.01	61.09	97.01	96.95	**1.488**	**0.927**	225.79	2019
ICNet [34]	ResNet-50	63.40	73.04	100.0	100.0	47.859	15.426	138.55	2018
UperNet [29]	ResNet-50	70.67	80.32	100.0	100.0	66.927	237.0	85.06	2018
BiSeNetV2 [33]	BiSeNetV2	66.53	76.32	98.08	98.28	14.821	12.286	172.48	2021
DDRNet [15]	DDRNet	68.91	78.65	99.94	99.94	5.766	4.56	156.38	2022
RepViT [24]	RepViT-M0.9	68.03	77.93	100.0	100.0	8.954	25.404	84.30	2024
GasTwinFormer	**MixTwinEncoder**	**74.47**	**83.63**	**100.0**	**100.0**	3.348	3.428	114.9	2025

Table 2. Comparison with state-of-the-art methods on our beef cattle methane emission dataset. ↑ indicates higher is better, ↓ indicates lower is better. Bold indicates better

polynomial decay (power=1.0). For GasTwinFormer, we initialize compatible components (patch embeddings, efficient attention, feed-forward networks) from SegFormer pre-trained weights while LSA layers are randomly initialized due to architectural novelty, using $10\times$ learning rate scaling for the decoder head and zero weight decay for normalization layers. Input images are resized to 512×512 pixels with data augmentation including random horizontal flipping (50% probability) and photometric distortion, using batch size 8 for training, and batch size 1 for inference. The multi-task pipeline handles simultaneous segmentation and classification annotations with validation every 8,000 iterations, retaining the top 3 checkpoints based on mean IoU performance. We report segmentation performance using mean Intersection over Union (mIoU) and mean F1-score (mF1), classification performance using accuracy and F1-score, and computational efficiency via parameters, FLOPs, and inference speed (FPS), with all metrics computed on the test set using the best validation checkpoint.

5.2. Comparison with state-of-the-arts

We evaluate GasTwinFormer against transformer-based and CNN-based methods on our beef cattle methane emission dataset. Table 2 summarizes performance metrics for segmentation and dietary classification tasks.

Segmentation Performance Analysis. GasTwinFormer achieves 74.47% mIoU and 83.63% mF1 using only 3.348M parameters and 3.428G FLOPs, outperforming all other approaches in terms of accuracy while maintaining exceptional efficiency. For instance, compared to Gasformer, GasTwinFormer delivers 2.22% better mIoU while requiring 9.9% fewer parameters and 65.4% fewer FLOPs. Compared to SegFormer, GasTwinFormer achieves 2.36% better mIoU and 2.06% better mF1 while requiring 11.5% fewer parameters and 56.5% fewer FLOPs. Moreover, GasTwinFormer outperforms all transformer-based approaches, including Twins PCPVT-s, achieving 0.42% better mIoU while being significantly more efficient with $8.3\times$ fewer parameters and $12.9\times$ fewer FLOPs. Compared to heavyweight CNN methods, our results demonstrate substantial superiority. Our method represents a 3.8% improvement over UperNet and a 4.11% improvement over DeepLabV3, while requiring $20\times$ fewer parameters and running 69–$78\times$ more efficiently in terms of FLOPs. Among efficient approaches, GasTwinFormer significantly outperforms Fast-FCNN by 20.46% mIoU and DDRNet by 5.56% mIoU.

GasTwinFormer delivers exceptional inference speed of 114.9 FPS, enabling real-time processing for practical livestock monitoring applications. Our method runs $1.87\times$ faster than Twins PCPVT-s and $2.23\times$ faster than Twins SVT-s, while also outperforming RepViT ($1.36\times$ faster), UperNet ($1.35\times$ faster), and DeepLabV3 ($1.25\times$ faster). Notably, while SegFormer achieves slightly higher FPS (119.66), GasTwinFormer delivers superior accuracy with 2.36% better mIoU. Compared to efficient CNN architectures, our method maintains competitive speed while delivering substantially better accuracy: it runs $2.11\times$ slower than BiSeNetV1 but achieves 21.6% better mIoU.

Dietary Classification Performance. GasTwinFormer achieves perfect dietary classification accuracy of 100% across all test samples, matching the performance of several state-of-the-art methods including Gasformer, SegFormer, and Twins variants. This demonstrates that our architectural design preserves multi-task learning capability while optimizing segmentation performance. Compared to methods with degraded classification performance, GasTwinFormer outperforms Fast-FCNN by 2.99%, BiSeNetV1 by 4.68%, and BiSeNetV2 by 1.92%, confirming the effectiveness of our Stage 4 feature extraction strategy for capturing dietary-specific emission patterns.

Qualitative Comparison. Figure 2 demonstrates distinct performance patterns across methods and dietary treat-

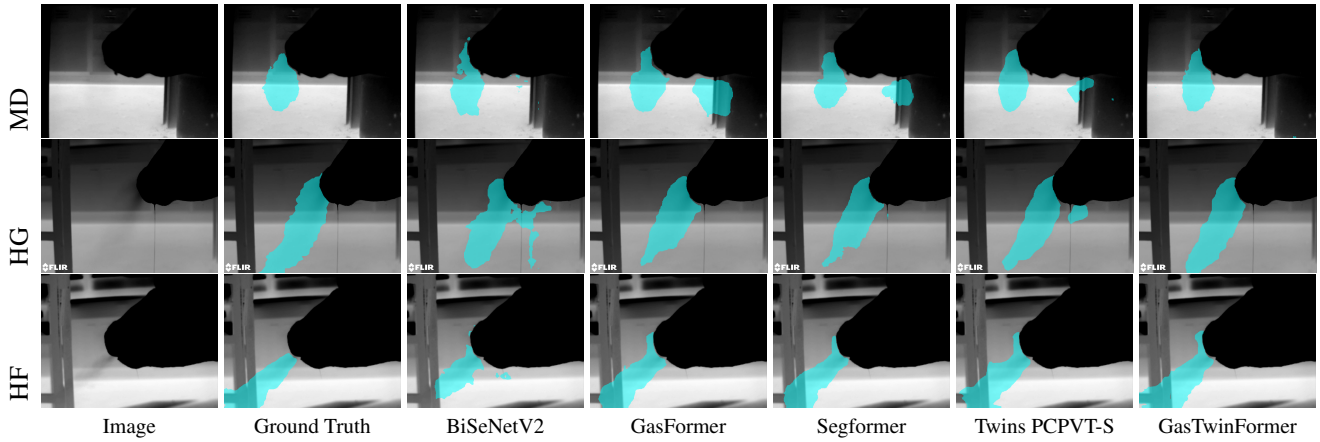

Figure 2. Qualitative comparison of methane plume segmentation results across different models and dietary treatments (MD: mixed diet, HG: high grain, HF: high forage), with ground truth masks shown for reference.

ments. CNN-based BiSeNetV2 consistently produces fragmented predictions with noise artifacts. Transformer methods (GasFormer, SegFormer, Twins PCPVT-S) generate false positives, predicting gas emissions in regions where ground truth shows none, particularly evident in the MD scenarios, while suffering incomplete coverage in HG and HF treatments, with additional over-segmentation in HF cases. GasTwinFormer maintains accurate delineation in MD, comprehensive coverage in both HG and HF cases despite minor over-segmentation in the latter case.

5.3. Ablation Studies

We evaluate each component of our GasTwinFormer architecture to validate design decisions. Unless specified, all ablations use 7×7 LSA windows, Mix-FFN, LRASPP decoder with F1+F2+F3 branch features, F4 for classification head input, Cross Entropy loss for both tasks, 128 internal decoder channels for segmentation, 256 hidden channels for classification, and LE attention pattern per stage.

Decoder head architecture comparison. We evaluate five different decoder heads to determine the best architecture for methane segmentation. Each decoder is configured with its standard design parameters reported in literature while maintaining consistent backbone features. As shown in Tab. 3, the LR-ASPP decoder achieves the best performance at 73.65% mIoU while maintaining optimal efficiency. Although ISA [6] and ANN [37] heads require significantly more parameters, they deliver inferior performance, validating our lightweight design approach.

Mix-FFN vs. Regular FFN. We compare Mix-FFN against standard feed-forward networks within LSA blocks. Table 3 demonstrates that Mix-FFN substantially outperforms regular FFN, achieving a 1.58 percentage point improvement. Mix-FFN incorporates spatial inductive bias through 3×3 depth-wise convolutions, crucial for capturing local spatial relationships without explicit positional encodings.

LR-ASPP Multi-scale feature fusion evaluation. We systematically evaluate different encoder feature combinations for the branch pathway of our Adaptive LR-ASPP decoder. We test all possible combinations of F1, F2, and F3 features to determine the optimal multi-scale fusion strategy. Table 3 compares performance across all branch combinations. F1+F2+F3 fusion delivers optimal performance, validating our design choice to utilize features from the first three encoder outputs for multi-scale branch processing. Notably, F2+F3 provides competitive performance with reduced computational cost, while individual feature configurations consistently underperform.

LR-ASPP decoder channel analysis. We examine how the decoder's internal channel dimension affects LR-ASPP performance. Table 4 shows performance, FLOPs, and parameters across different channel configurations. Our experiments demonstrate that 128 decoder channels deliver the best segmentation accuracy while maintaining computational efficiency. Increasing channels beyond 128 decreases performance while dramatically increasing computational overhead. For example, 1024 channels achieves the second-best mIoU but still underperforms 128 channels while requiring $2.4\times$ more parameters and $7.7\times$ more FLOPs.

Classification feature source evaluation. We examine which encoder stage provides optimal features for dietary classification. Table 4 compares performance across different encoder stage selections. Stage 4 features achieve the best segmentation performance and perfect classification accuracy, validating our design choice. In contrast, Stage 2 and Stage 3 show progressively reduced performance despite requiring fewer parameters.

Hybrid attention pattern evaluation. We systematically test different combinations of locally-grouped self-attention (L) and efficient multi-head attention (E) to identify the optimal hybrid pattern. Table 4 compares results across six attention configurations. EL-EL-EL-EL pattern achieves the highest performance, slightly outperforming our initial LE-LE-LE-LE baseline. Pure attention patterns demonstrate in-

Configuration	mIoU (%)↑	Diet Acc. (%)↑	Params (M)↓	FLOPs (G)↓
Decoder Head Types				
All-MLP [30]	73.42	99.94	3.548	7.591
FCN Head [11]	71.23	99.94	3.416	7.303
ISA Head [6]	70.02	**100.0**	4.666	**3.337**
ANN Head [37]	70.67	**100.0**	5.599	3.695
LR-ASPP [5]	**73.65**	**100.0**	**3.348**	3.508
LSA Feed-Forward Network				
Regular FFN	72.07	**100.0**	**3.065**	**3.471**
MixFFN	**73.65**	**100.0**	3.085	3.508
LR-ASPP Multi-Scale Feature Fusion				
F1 only	72.70	**100.0**	**3.256**	3.323
F1+F2	72.39	**100.0**	3.285	3.443
F2 only	72.55	**100.0**	3.261	3.407
F2+F3	72.78	**100.0**	3.326	3.140
F1+F3	72.60	**100.0**	3.319	3.387
F3 only	72.34	**100.0**	3.297	**3.019**
F1+F2+F3	**73.65**	**100.0**	3.348	3.508

Table 3. Foundation architecture component studies establishing core design choices through systematic optimization of decoder head, LSA feed-forward network, and multi-scale feature fusion.

ferior performance, particularly all local attention configurations. This validates that hybrid attention design is crucial, where efficient attention captures global context first, followed by local attention refinement.

Loss function comparison. We evaluate the effectiveness of Gaussian Plume Weighted Dice Loss [36] against standard segmentation losses including Cross Entropy, Dice, and Focal loss for segmentation, while maintaining Cross Entropy for classification. Table 5 shows the results for this comparison. Gaussian Plume loss achieves the highest performance at 73.97% mIoU, outperforming all traditional loss functions. While Dice loss achieves perfect classification accuracy, its segmentation performance lags behind by 1.4 percentage points. Focal loss demonstrates the poorest performance across both tasks. This demonstrates that domain-specific physical modeling particularly benefits the segmentation task by leveraging the inherent characteristics of gas plume dynamics.

Configuration	mIoU (%)↑	Diet Acc. (%)↑	Params (M)↓	FLOPs (G)↓
LR-ASPP Channel Scaling				
128 ch	**73.65**	**100.0**	**3.348**	**3.508**
256 ch	72.32	**100.0**	3.644	4.729
512 ch	72.37	**100.0**	4.630	9.309
768 ch	72.97	99.38	6.140	16.741
1024 ch	73.03	99.77	8.174	27.025
2048 ch	72.72	**100.0**	21.555	96.684
Classification Feature Source				
Encoder Stage 4	**73.65**	**100.0**	3.348	**3.508**
Encoder Stage 3	69.52	**100.0**	3.323	3.508
Encoder Stage 2	71.91	**100.0**	**3.299**	3.508
Hybrid Attention Pattern Analysis†				
LE-LE-LE-LE	73.65	**100.0**	3.348	3.508
EL-EL-EL-EL	**73.69**	98.70	3.348	3.508
LL-LL-LL-LL	68.97	98.59	**3.113**	**3.214**
EE-EE-EE-EE	73.17	**100.0**	3.582	3.802
LL-LL-EE-EE	70.56	99.77	3.319	3.259
EE-EE-LL-LL	73.60	**100.0**	3.376	3.757

Table 4. Architecture refinement and pattern optimization studies including decoder channel scaling, classification feature source selection, and systematic evaluation of hybrid attention patterns.

Configuration	mIoU (%)↑	Diet Acc. (%)↑	Params (M)↓	FLOPs (G)↓
Loss Function Comparison				
Cross Entropy Loss	73.69	98.70	3.348	3.508
Dice Loss	72.57	**100.0**	3.348	3.508
Focal Loss	70.33	98.65	3.348	3.508
Gaussian Plume Loss	**73.97**	99.44	3.348	3.508
LSA Window Size Optimization				
7×7	73.97	99.44	3.348	3.508
5×5	**74.47**	**100.0**	3.348	3.428
3×3	74.35	**100.0**	3.348	**3.367**

Table 5. Task-specific loss and parameter optimization studies comparing segmentation losses and LSA window size refinement.

LSA window size optimization. Finally, we analyze the influence of LSA window size using our best configuration with EL-EL-EL-EL pattern and Gaussian Plume loss. Table 5 compares performance and efficiency across different window sizes. Our analysis reveals that 5×5 windows achieve the highest performance at 74.47% mIoU, outperforming both 3×3 and baseline 7×7 windows. The 5×5 size provides optimal balance between local receptive field coverage and computational efficiency. Moderate window sizes prove most effective for capturing gas plume local structures.

6. Conclusion

We presented GasTwinFormer, a hybrid vision transformer for livestock methane emission segmentation and dietary classification. Comprehensive benchmarking on our beef cattle methane dataset demonstrates that GasTwinFormer outperforms all state-of-the-art methods, achieving superior segmentation and dietary classification performance with significantly fewer computational requirements. Extensive ablation studies validate our architectural design choices. While our current study focuses on beef cattle, the architecture is directly extensible to other ruminant species in free-range grazing and broader gas detection contexts by fine-tuning window sizes and training on species-specific OGI data. This work establishes a strong foundation for automated livestock emission monitoring and climate mitigation applications.

Acknowledgement

This work is supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number 2022-70001-37404, and by the Office of the Vice Chancellor for Research at Southern Illinois University Carbondale.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5, 6

[2] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in neural information processing systems*, 34:9355–9366, 2021. 1, 2, 3, 5, 6

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[4] Mohamed G Embaby, Toqi Tahamid Sarker, Amer AbuGhazaleh, and Khaled R Ahmed. Optical gas imaging and deep learning for quantifying enteric methane emissions from rumen fermentation in vitro. *IET Image Processing*, 19(1):e13327, 2025. 1

[5] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 1, 3, 4, 8

[6] Lang Huang, Yuhui Yuan, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Interlaced sparse self-attention for semantic segmentation. *arXiv preprint arXiv:1907.12273*, 2019. 1, 7, 8

[7] Ismot Jahan, Mohamed Mehana, Georgios Matheou, and Hari Viswanathan. Deep learning-based quantifications of methane emissions with field applications. *International Journal of Applied Earth Observation and Geoinformation*, 132:104018, 2024. 1, 2

[8] Jinpeng Li, Yichao Yan, Shengcai Liao, Xiaokang Yang, and Ling Shao. Local-to-global self-attention in vision transformers. *arXiv preprint arXiv:2107.04735*, 2021. 1

[9] Weicong Liang, Yuhui Yuan, Henghui Ding, Xiao Luo, Weihong Lin, Ding Jia, Zheng Zhang, Chao Zhang, and Han Hu. Expediting large-scale vision transformer for dense prediction without fine-tuning. *Advances in Neural Information Processing Systems*, 35:35462–35477, 2022. 1

[10] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 8

[12] Office of Laboratory Animal Welfare. Institutional animal care and use committee guidebook. Guidebook, National Institutes of Health, Bethesda, MD, 2002. 5

[13] Seán O'Connor, Flannagán Noonan, Desmond Savage, and Joseph Walsh. Advancements in real-time monitoring of enteric methane emissions from ruminants. *Agriculture*, 14(7): 1096, 2024. 1

[14] Rajendra K Pachauri, Myles R Allen, Vicente R Barros, John Broome, Wolfgang Cramer, Renate Christ, John A Church, Leon Clarke, Qin Dahe, Purnamita Dasgupta, et al. *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Ipcc, 2014. 1

[15] Huihui Pan, Yuanduo Hong, Weichao Sun, and Yisong Jia. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 5, 6

[16] Rudra PK Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-scnn: Fast semantic segmentation network. *arXiv preprint arXiv:1902.04502*, 2019. 5, 6

[17] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 2

[18] Bertrand Rouet-Leduc and Claudia Hulbert. Automatic detection of methane emissions in multispectral satellite imagery using a vision transformer. *Nature Communications*, 15(1):3801, 2024. 2

[19] Toqi Tahamid Sarker, Mohamed G Embaby, Khaled R Ahmed, and Amer AbuGhazaleh. Gasformer: A transformer-based architecture for segmenting methane emissions from livestock in optical gas imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5489–5497, 2024. 2, 5, 6

[20] Tim Searchinger, Richard Waite, Craig Hanson, Janet Ranganathan, Patrice Dumas, and Emily Matthews. Creating a sustainable food future: a menu of solutions to feed nearly 10 billion people by 2050-synthesis report. 2018. 1

[21] Drew Shindell, AR Ravishankara, Johan CI Kuylenstierna, Eleni Michalopoulou, Lena Höglund-Isaksson, Yuqiang Zhang, Karl Seltzer, Muye Ru, Rithik Castelino, Greg Faluvegi, et al. Global methane assessment: Benefits and costs of mitigating methane emissions. Technical report, United Nations Environment Programme, 2021. 1

[22] Luis Orlindo Tedeschi, Adibe Luiz Abdalla, Clementina Alvarez, Samuel Weniga Anuga, Jacobo Arango, Karen A Beauchemin, Philippe Becquet, Alexandre Berndt, Robert Burns, Camillo De Camillis, et al. Quantification of methane emitted by ruminants: a review of methods. *Journal of Animal Science*, 100(7):skac197, 2022. 1

[23] Teledyne FLIR. *FLIR G-Series: Gx320, G620, Gx620 Optical Gas Imaging (OGI) Cameras for Hydrocarbons*. Teledyne FLIR, LLC, 2023. Datasheet. 4

[24] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15909–15920, 2024. 5, 6

[25] Jingfan Wang, Lyne P Tchapmi, Arvind P Ravikumar, Mike McGuire, Clay S Bell, Daniel Zimmerle, Silvio Savarese,

and Adam R Brandt. Machine vision for natural gas methane emissions detection using an infrared camera. *Applied Energy*, 257:113998, 2020. 1, 2

[26] Jingfan Wang, Jingwei Ji, Arvind P Ravikumar, Silvio Savarese, and Adam R Brandt. Videogasnet: Deep learning for natural gas methane leak classification using an infrared camera. *Energy*, 238:121516, 2022. 2

[27] Jiayang Lyra Wang, Brenna Barlow, Wes Funk, Cooper Robinson, Adam Brandt, and Arvind P Ravikumar. Large-scale controlled experiment demonstrates effectiveness of methane leak detection and repair programs at oil and gas facilities. *Environmental Science & Technology*, 58(7):3194–3204, 2024. 1

[28] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 2

[29] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 5, 6

[30] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 1, 2, 3, 4, 5, 6, 8

[31] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE/CVF International Conference on Computer vision*, pages 16269–16279, 2021. 1, 2

[32] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 5, 6

[33] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, pages 1–18, 2021. 5, 6

[34] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 405–420, 2018. 5, 6

[35] Chuanyang Zheng. iformer: Integrating convnet and transformer for mobile application. *arXiv preprint arXiv:2501.15369*, 2025. 5, 6

[36] Jiani Zhou, Yang Liu, Yong Zhang, Haotian Hu, Zenan Leng, Feng Sun, and Chen Chen. High-accuracy combustible gas cloud imaging system using yolo-plume classification network. *Frontiers in Physics*, 13:1603047, 2025. 4, 8

[37] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 593–602, 2019. 7, 8