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

Anonymous ICCV submission

Paper ID *****

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.

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 [21, 22]. Agriculture accounts for 40% of human-caused methane emissions, with livestock responsible for roughly 32% [3]. 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 [21].

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 [5, 7]. This biological relationship enables simultaneous detection of methane emissions and classification of dietary treatments, allowing farmers to implement evidence-based feeding strategies for emission reduction [6].

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 [20]. Recent advances in optical gas imaging (OGI) offer non-invasive, continuous monitoring capabilities using thermal infrared cameras operating in the 3.2-3.4 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 [10].

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 [11]. Recent hybrid attention mechanisms show promise for balancing efficiency with representational capacity, but have not been adapted for gas plume segmentation [23, 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 [30] with locally-grouped self-attention [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 [8] that processes features from multiple encoder stages to generate accurate segmentation predictions. Our framework performs both pixel-wise methane segmentation and dietary classification using 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 manually 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 99% 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 [17] and CNNs for airborne emission quantification [10]. 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.* [4] established Vision Transformers for image classification, while Ranftl *et al.* [16] introduced Dense Vision Transformers for dense prediction tasks. Hierarchical designs have proven effective: Swin Transformer [12] uses shifted windowing for computational efficiency, PVT [28] establishes hierarchical principles through progressive spatial reduction, and SegFormer [30] achieves state-of-the-art performance (49.02% 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 [8] for multi-scale feature aggregation, and (3) a dietary classification head for pixel-wise 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 dimensions. The alternating pattern operates as follows: the EMA Block first establishes global relationships through spatially-reduced attention, then the LSA refines local structures through windowed attention within 5×5 local regions. Each stage contains one EMA-LSA block pair (denoted as EL), where the EMA block precedes the LSA block within the same stage.

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. Each stage contains one EMA-LSA block pair with overlapped patch embedding and layer normalization, totaling 8 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 attention mechanism from SegFormer [1],

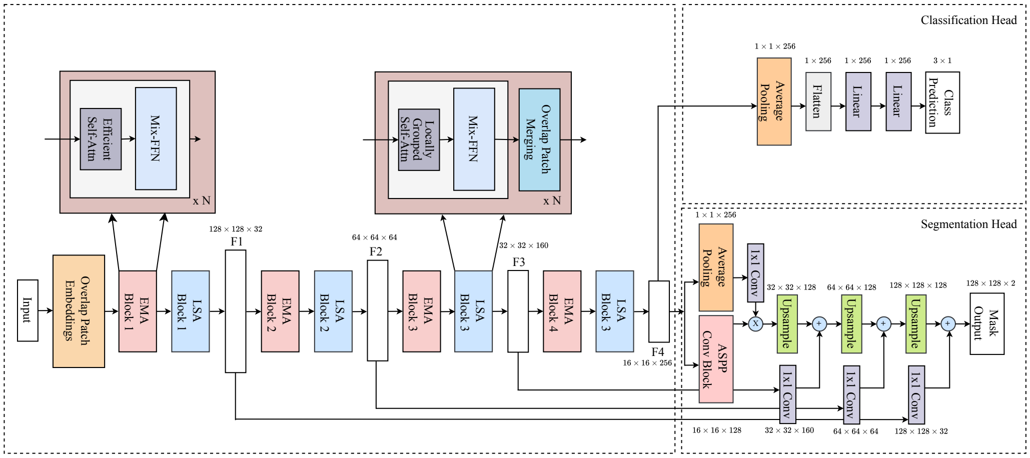

Figure 1. Overview of GasTwinFormer architecture. The Mix Twin encoder alternates between Efficient Multi-head Attention (EMA) and Locally-grouped Self-Attention (LSA) blocks across four hierarchical stages, generating multi-scale features $\{F_1, F_2, F_3, F_4\}$ at resolutions $\{H/4 \times W/4, H/8 \times W/8, H/16 \times W/16, H/32 \times W/32\}$. The lightweight LR-ASPP decoder aggregates these features for dense prediction, while dual task heads enable simultaneous methane segmentation and dietary classification from stage 4 features.

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.

This approach implements spatial reduction on key and value representations through strided convolution operations. 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 efficient multi-head attention 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 Locally-Grouped Self-Attention (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 mechanism 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 $w_1 = w_2 = 7$, the complexity becomes $\mathcal{O}(49HWd)$, 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:

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

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 such as semantic segmentation, we use a lightweight decoder design that efficiently aggregates multi-scale features from our hierarchical encoder. Building upon the Lite Reduced Atrous Spatial Pyramid Pooling (LR-ASPP) decoder introduced in MobileNetV3 [8], we propose an adaptive variant that accommodates variable input resolutions and backbone architectures while maintaining computational efficiency.

The standard R-ASPP decoder from MobileNetV2 [18] employs two branches consisting of a 1×1 convolution and global average pooling operation. However, LR-ASPP improves upon this design by deploying global average pooling in a fashion similar to the Squeeze-and-Excitation module [9], employing adaptive pooling with simplified channel processing to reduce computational overhead.

Our Hierarchical LR-ASPP decoder processes multi-scale features $\{F_1, F_2, F_3, F_4\}$ from the hierarchical encoder through the following operations:

$$\begin{aligned} F_{\text{pool}} &= \text{Sigmoid}(\text{Conv}_{1\times 1}(\text{AdaptiveAvgPool}(F_4))) \\ F_{\text{aspp}} &= \text{Conv}_{1\times 1}(F_4) \odot \text{Upsample}(F_{\text{pool}}) \\ F_{\text{out}} &= \text{ProgressiveUpsampling}(F_{\text{aspp}}, \{F_3, F_2, F_1\}) \end{aligned} \tag{5}$$

where $\odot$ denotes element-wise multiplication, and the adaptive average pooling operation replaces fixed-size kernels to ensure compatibility with variable input resolutions. The decoder progressively aggregates features through skip connections and upsampling operations, preserving both semantic information from deep features and spatial details from shallow features essential for accurate methane plume boundary delineation. The lightweight design achieves linear computational scaling $\mathcal{O}(C \times H \times W)$ with respect to spatial resolution, making it suitable for efficient deployment.

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), providing complementary information to the pixel-level segmentation predictions by associating methane emission patterns with specific feeding regimens.

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:

$$w(x, y) = \exp\left(-\frac{(x - \mu_x)^2}{2\sigma_x^2} - \frac{(y - \mu_y)^2}{2\sigma_y^2}\right) \tag{6}$$

where (μ_x, μ_y) represents the plume center computed via center-of-mass on predicted masks, and (σ_x, σ_y) denote the diffusion scales estimated through weighted standard deviation with adaptive bounds $[\frac{W}{20}, \frac{W}{2}]$ and $[\frac{H}{20}, \frac{H}{2}]$ to ensure numerical stability. The weighted Dice loss is then computed as:

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

4. Beef Cattle Methane Emission Dataset

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, 30 fps recording rate, and $<$15 mK thermal sensitivity. The camera detects methane concentrations as low as 200 ppm·m under optimal conditions with minimum 2–3°C thermal contrast.

We present a comprehensive dataset for methane emission detection from dairy 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.

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 (1200lb±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 30lb of diet mix per cow daily. All cows received 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 barns, managed in accordance with Institutional Animal Care and Use Committee guidelines (protocol number 22-016) [13]. 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 where gas plumes appear as dark regions against lighter backgrounds. Raw thermal data was recorded in FLIR's CSQ format preserving 14-bit radiometric information. 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%) containing visible methane plumes, reflecting the intermittent nature of bovine eructation events. 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. The resulting splits contain 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 classical processing stage 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 using vertical and elliptical kernels, with size filtering retaining regions >2000 pixels and eccentricity filtering (>0.95) removing linear artifacts. Using initial masks as training data, we trained a Gasformer [19] transformer-based model adapted for single-class plume segmentation to identify subtle patterns beyond classical methods. Our enhanced processing pipeline applies CLAHE, intensity rescaling, and non-local means denoising ($h = 15$) to generate improved candidate masks. For each frame, we compare three mask candidates from classical processing, Gasformer predictions, and enhanced processing, with manual inspection selecting the most accurate representation based on visual assessment using contrast-enhanced overlays, leveraging complementary strengths across methods while addressing inherent uncertainty in gas plume boundaries.

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], Gasformer [19], RepViT [24], iFormer [35], Twins PCPVT-S [2], Twins SVT-S [2]) and CNN-based methods (Fast-FCNN [15], BiSeNetV1 [32], BiSeNetV2 [33], DDRNet [14], ICNet [34], UperNet [29], DeepLabV3[1]), with all models utilizing ImageNet pre-trained 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 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× learning rate scaling for the decoder head and zero weight decay for nor-

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

Table 1. 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

malization 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 compare our results with existing approaches on our beef cattle methane emission dataset. Table 1 summarizes our results including parameters, FLOPs, inference speed, and accuracy for both segmentation and dietary classification tasks. We organize methods into transformer-based and CNN-based approaches to provide comprehensive comparisons across different architectural paradigms.

Segmentation Performance Analysis. As shown on our methane emission dataset, 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× fewer parameters and 12.9× 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× fewer parameters and running 69–79× more efficiently in terms of FLOPs. Among efficient approaches, GasTwinFormer significantly outperforms Fast-FCNN by 20.46% mIoU and DDRNet by 5.56% mIoU, establishing clear superiority in the accuracy-efficiency trade-off space.

GasTwinFormer delivers exceptional inference speed of 114.9 FPS, enabling real-time processing for practical livestock monitoring applications. Our method runs 1.87× faster than Twins PCPVT-s and 2.23× faster than Twins SVT-s, while also outperforming RepViT (1.36× faster), UperNet (1.35× faster), and DeepLabV3 (1.25× 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× 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 clear

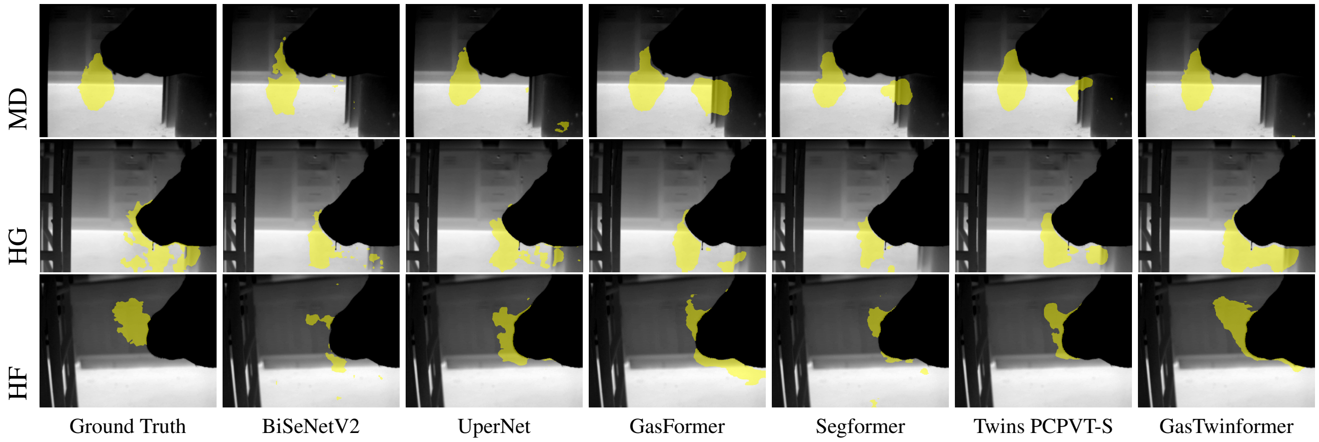

Figure 2. Qualitative comparison of methane plume segmentation results across different models and dietary treatments (MD: mixed diet, HG: high grain, HF: high forage), with ground truth masks shown for reference.

qualitative differences across methods and dietary treatments. CNN-based approaches (BiSeNetV2, UperNet) produce fragmented predictions with significant noise artifacts. Transformer methods (Gasformer, SegFormer, Twins PCPVT-S) tend to generate false positives, predicting gas emissions in regions where ground truth shows none, particularly evident in the MD scenarios. In HG cases with fragmented ground truth patterns, GasTwinFormer occasionally over-masks by connecting scattered emission regions.

5.3. Ablation Studies

We systematically evaluate each component of our GasTwinFormer architecture to validate design decisions. Unless otherwise specified, all experiments use Cross Entropy loss for both segmentation and dietary classification tasks.

Decoder head architecture comparison. We first evaluate five different decoder heads to determine the optimal architecture for methane segmentation. Table 2 presents results using our baseline configuration with LE-LE-LE-LE attention pattern, S1+S2+S3 multi-scale features, 128 channels, and stage 4 classification. The LR-ASPP decoder achieves the best performance at 73.65% mIoU while maintaining optimal efficiency. Although ISA and ANN heads require significantly more parameters, they deliver inferior performance, validating our lightweight design approach. These findings establish LR-ASPP as our decoder choice due to its superior accuracy-efficiency characteristics.

Mix-FFN vs. Regular FFN. We compare Mix-FFN against standard feed-forward networks within LSA blocks. Table 2 demonstrates that Mix-FFN substantially outperforms regular FFN, achieving a 1.58 percentage point improvement. The Mix-FFN incorporates spatial inductive bias through 3×3 depth-wise convolutions, which proves crucial for capturing local spatial relationships in gas plume boundaries without requiring explicit positional encodings. This validates that Mix-FFN enhances feature representations for dense prediction tasks.

Multi-scale feature fusion evaluation. We systematically test different encoder stage combinations for multi-scale feature fusion. Table 2 compares performance across all combinations of stages S1, S2, and S3. S1+S2+S3 fusion delivers optimal performance, validating our design choice. Notably, S2+S3 provides competitive performance with reduced computational cost, while individual stages consistently underperform. This confirms that multi-scale fusion is essential for accurate gas plume segmentation by enabling capture of both fine-grained boundary details and semantic context.

LR-ASPP channel dimension analysis. We examine how channel dimension C affects LR-ASPP decoder performance. Table 3 shows performance, FLOPs, and parameters across different channel configurations. Our experiments demonstrate that 128 channels deliver the best segmentation accuracy while maintaining computational efficiency. Increasing channels beyond 128 generally decreases performance while dramatically increasing computational overhead. For example, even the best alternative at 1024 channels still underperforms 128 channels while requiring 2.4× more parameters and 7.7× more FLOPs. Based on this analysis, we select 128 channels for our decoder to achieve optimal balance between accuracy and efficiency.

Classification feature source evaluation. We examine which encoder stage provides optimal features for dietary classification. Table 3 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. This indicates that deeper, more semantic features from Stage 4 are crucial for both accurate segmentation and reliable dietary classification. High-level semantic features prove essential for our

Configuration	mIoU (%)↑	Diet Acc. (%)↑	Params (M)↓	GFLOPs ↓
Decoder Head Types				
SegFormer Head	73.42	99.94	3.548	7.591
FCN Head	71.23	99.94	3.416	7.303
ISA Head	70.02	100.0	4.666	3.337
ANN Head	70.67	100.0	5.599	3.695
LR-ASPP	**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
Multi-Scale Feature Fusion				
S1 only	72.70	100.0	**3.256**	3.323
S1+S2	72.39	100.0	3.285	3.443
S2 only	72.55	100.0	3.261	3.407
S2+S3	72.78	100.0	3.326	3.140
S1+S3	72.60	100.0	3.319	3.387
S3 only	72.34	100.0	3.297	**3.019**
S1+S2+S3	**73.65**	**100.0**	3.348	3.508

All models trained with LE-LE-LE-LE pattern & Cross Entropy loss
Fixed: S1+S2+S3, 128ch, Stage 4 classification (unless ablated)

Table 2. Foundation architecture component studies establishing core design choices through systematic optimization of decoder head, LSA feed-forward network, and multi-scale feature fusion using LE-LE-LE-LE baseline configuration.

Configuration	mIoU (%)↑	Diet Acc. (%)↑	Params (M)↓	GFLOPs ↓
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

Channel scaling & classification studies: LE-LE-LE-LE + Cross Entropy
† Attention pattern study: Cross Entropy loss, pattern varies

Table 3. Architecture refinement and pattern optimization studies including decoder channel scaling, classification feature source selection, and systematic evaluation of hybrid attention patterns to identify optimal EL-EL-EL-EL configuration.

Configuration	mIoU (%)↑	Diet Acc. (%)↑	Params (M)↓	GFLOPs ↓
Loss Function Comparison				
Cross Entropy	73.69	98.70	3.348	3.508
Gaussian Plume	**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 4. Task-specific loss and parameter optimization studies comparing Cross Entropy vs Gaussian Plume loss across model configurations, and final LSA window size refinement using optimized Gaussian Plume loss and EL-EL-EL-EL attention pattern.

dual-task architecture.

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 3 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 inferior 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 Cross Entropy loss for segmentation, while maintaining Cross Entropy for classification. Table 4 shows the results for this comparison. Gaussian Plume loss delivers improved performance over Cross Entropy baseline. Interestingly, improved segmentation performance also improves classification accuracy despite unchanged classification loss. This demonstrates that incorporating physical gas dispersion constraints through Gaussian weighting benefits the shared encoder-decoder architecture.

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 4 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. This work establishes a strong foundation for automated livestock emission monitoring and climate mitigation applications.

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

[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

[3] Clean Air Task Force. Accelerating climate solutions in agriculture: Why reducing methane from livestock is an urgent opportunity. https://www.catf.us/2024/10/accelerating-climate-solutions-agriculture-why-reducing-methane-livestock-urgent-opportunity/, 2024. 1

[4] 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

[5] 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

[6] Food Forward NDCs. Reducing emissions from livestock through sustainable management practices. https://foodforwardndcs.panda.org/food-production/reducing-emissions-from-livestock-through-sustainable-management-practices/, 2024. 1

[7] Jalil Ghassemi Nejad, Mun-Su Ju, Jang-Hoon Jo, Kyung-Hwan Oh, Yoon-Seok Lee, Sung-Dae Lee, Eun-Joong Kim, Sanggun Roh, and Hong-Gu Lee. Advances in methane emission estimation in livestock: A review of data collection methods, model development and the role of ai technologies. *Animals*, 14(3):435, 2024. 1

[8] 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, 2, 4

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[10] 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

[11] 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

[12] 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

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

[14] 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

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

[16] 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

[17] 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

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 4

[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

[20] 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

[21] United Nations Environment Programme. Methane emissions are driving climate change. here's how to reduce them. https://www.unep.org/news-and-stories/story/methane-emissions-are-driving-climate-change-heres-how-reduce-them, 2024. 1

[22] University of Wisconsin Extension. Methane emissions from livestock and climate change. https://cropsandsoils.extension.wisc.edu/articles/methane-emissions-from-livestock-and-climate-change/, 2024. 1

[23] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceed-*

ings of the IEEE/CVF international conference on computer vision, pages 5785–5795, 2023. 1

[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

[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

[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, 5

[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

[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

[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

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

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