ANCHOR-VIT: SPATIALLY-FOCUSED VISION TRANSFORMER FOR DISTRACTED DRIVING DETECTION

Vivan Doshi
Bellarmine College Preparatory
Email: vivandoshi24@gmail.com

***Abstract*—Distracted driving remains a critical threat to road safety, often leading to severe accidents and fatalities. Traditional driver monitoring solutions, including CNNs and standard Vision Transformers, frequently fail to capture subtle, spatially localized cues that signal driver distraction. This paper presents Anchor-ViT, a novel Vision Transformer architecture that integrates learnable spatial anchors with a Soft Radial Attention (SRA) mechanism to adaptively focus on driver-critical areas. These anchors are optimized via gradient descent to guide attention toward relevant patches, while SRA employs a Gaussian kernel to reinforce local interactions and preserve global context through a dedicated class token. Evaluations on the State Farm and 100-Driver distracted driving datasets show that Anchor-ViT outperforms baseline ViT models by up to 5.2% in accuracy, effectively balancing the need for localized sensitivity and comprehensive scene understanding. This innovative design holds promise for enhancing driver monitoring, improving overall road and driver safety.**

***Index Terms*—Distracted Driving Detection, Vision Transformer, Driver Monitoring, Deep Learning**

I. INTRODUCTION

Driver distraction is a major threat to global road safety, significantly increasing traffic accidents and resulting in countless injuries and fatalities. According to the NHTSA, in the United States during 2021, distracted driving contributed to 8 percent of fatal crashes, 14 percent of injury crashes, and 13 percent of all police-reported motor vehicle traffic crashes [1]. This led to the devastating loss of 3,522 lives and an estimated 362,415 injuries in the U.S. in 2021 alone [1]. These figures demonstrate the immense human and economic costs associated with distracted driving, emphasizing the urgent need for effective countermeasures [2]–[4]. While various approaches have been explored [5], [6], a persistent challenge remains in accurately and reliably detecting the diverse and often subtle aspects of driver distraction in real-world scenarios [2]–[4].

Current driver monitoring systems, even those leveraging deep learning methods, often fall short in capturing the full spectrum of driver distraction behaviors [2]–[4]. A core limitation lies in their ability to effectively process the visual information crucial for identifying distraction cues. While these systems can recognize broad scene context, they frequently struggle with the spatially localized nature of driver distraction. Consider subtle indicators: a downward glance at a phone, a hand movement towards the infotainment system, or minute shifts in gaze reflecting cognitive disengagement. These cues are concentrated in specific regions – around the hands, the driver's face, the dashboard, and interaction zones. Existing systems, designed for general-purpose image analysis, may miss these vital signals by distributing attention broadly across the entire scene. They lack mechanisms to prioritize and focus on key spatial areas where distraction is most evident. This hinders their ability to distinguish between genuinely distracted behaviors and normal driving actions, leading to missed detections or false alarms, and limiting their effectiveness in real-world applications. The crucial missing element in current solutions is a robust method to spatially focus attention within the driving scene, targeting driver-relevant regions to capture subtle distraction indicators.

To overcome these limitations and enhance distracted driving detection, this paper introduces a novel approach embodied in the Anchor-ViT architecture. Our method addresses spatially-aware distraction detection by equipping a Vision Transformer (ViT) framework [7] with a mechanism to intelligently focus on driver-relevant areas. Think of a standard Vision Transformer as illuminating the entire image with a floodlight, casting a broad light across everything. In contrast, Anchor-ViT operates like spotlights, directing its processing power precisely to the most relevant areas. Anchor-ViT achieves this through the integration of learnable spatial anchors and a Soft Radial Attention (SRA) mechanism. These anchors act as guideposts, strategically positioned within the image to direct attention towards regions like the face, hands, and interaction zones. The Soft Radial Attention (SRA) mechanism amplifies the model's focus around these guideposts by increasing the importance of visual information from their vicinity, while reducing the influence of distant, less relevant regions. For instance, when detecting a driver texting, Anchor-ViT's anchors might focus on the lower central region where hands and a phone appear. The SRA mechanism amplifies attention to this hand-phone region, allowing the model to effectively capture cues like hand movements and phone presence, compared to a standard ViT that might distribute attention uniformly across the dashboard and windshield. This targeted approach allows Anchor-ViT to prioritize critical cues of distraction without sacrificing the Vision Transformer's ability to understand the global driving context. By dynamically focusing on spatially relevant regions, Anchor-ViT becomes more sensitive to subtle cues, leading to a more accurate and robust detection system. This approach addresses the limitations of existing systems by ensuring that the most informative parts of the driving scene are prioritized in distraction detection.

Extensive evaluations on benchmark distracted driving datasets demonstrate that Anchor-ViT consistently outperforms conventional Vision Transformer models. This work

contributes in three ways. First, the Anchor-ViT architecture is introduced—a novel approach that enhances distracted driving detection by focusing on spatially relevant regions through learnable spatial anchors and a Soft Radial Attention mechanism. Second, comprehensive validation showcases improved accuracy and robustness compared to existing methods. Finally, detailed ablation studies demonstrate the contributions of each core component, solidifying the effectiveness of the approach. Collectively, these advancements improve driver monitoring systems, enhancing road safety and paving the way for more advanced driver assistance and autonomous vehicle technologies.

II. RELATED WORKS

Convolutional neural networks (CNNs) have become a cornerstone in distracted driving detection due to their ability to learn hierarchical representations directly from raw images. Early models were often designed with real-time detection in mind, employing lightweight architectures such as MobileVGG [8] to satisfy the computational constraints of in-vehicle systems [6]. Researchers further refined these models by introducing architectural modifications—such as decreasing filter sizes in CNNs to capture fine-grained spatial details—and developing efficient and lightweight CNN frameworks [9] to streamline real-time detection. To enhance the localization of subtle cues like a driver's hand movements or facial expressions, several approaches incorporated attention mechanisms, including those in improved YOLOv8 variants [10] and Bi-LSTM models [11]. Some methods integrated non-visual cues such as hand-grip sensing to complement visual data for faster detection [12]. Moreover, hybrid models that combine CNN-based local feature extraction with global contextual reasoning have demonstrated effectiveness [13], [14] in complex driving environments.

Building on these successes, subsequent research embraced transfer learning and ensemble strategies to achieve further performance improvements. Ensemble approaches combining transfer learned CNN architectures [15] were employed to enhance robustness and accuracy. Domain adaptation techniques have also been explored to improve generalization across different driving conditions and datasets [16]. Furthermore, contrastive learning approaches have been used for quantitative identification of driver distraction [17]. These advancements naturally paved the way for more sophisticated deep learning frameworks.

The most recent breakthroughs in the field have centered on Vision Transformers (ViTs) [7], which have gained prominence for their powerful self-attention mechanisms [18], [19] capable of modeling long-range dependencies and capturing global context. Efficient attention mechanisms for ViTs have also been a focus of research [20]. Early applications of ViTs in distracted driving detection demonstrated their effectiveness [21], [22]. Benchmarking studies have compared CNNs and ViTs for this task [23], and hybrid CNN-ViT models have been explored for efficient and lightweight detection [13], [24]. Subsequent research refined ViT-based models by incorporating transfer learning [22] and exploring different ViT architectures such as Shifted-Window Hierarchical Vision Transformers [25] and Swin Transformers [26], along with fine-grained detection using Feature Pyramid Vision Transformers [27]. Multi-task learning with Vision Transformers has also been used for distraction and emotion detection [28]. Multi-view and multi-scale Vision Transformers have been also proposed for improved driver action recognition [29], and spatio-temporal learning with transformers has been used for understanding driver behaviors in naturalistic videos [30]. Furthermore, video transformers have been applied for distracted driver recognition from temporal video data [31]. Attention mechanisms within transformers have also been augmented for naturalistic driving action recognition [32] and vision-language models like CLIP have also been explored to enhance robustness [33]. Explainability of driver activity recognition using video transformers has also been addressed [34]. Prompting techniques for guiding attention in ViTs [35] and unified local and global attention interaction modeling in ViTs [36] represent recent advancements in the field of vision transformers regarding its attention mechanisms. Contrastive learning with video transformers has been used for multi-view and multimodal video data [25], and unsupervised learning algorithms have been developed for fine-grained distraction detection [37].

In summary, the evolution from CNN-based frameworks—with their direct feature learning and real-time efficiency—through the integration of transfer learning and ensemble strategies, has set the stage for the transformative potential of Vision Transformers in distracted driving detection. By leveraging the global context modeling capabilities of ViTs and addressing their localization challenges through spatially guided attention, this approach represents a state-of-the-art solution in this critical application.

III. METHODOLOGY

This paper introduces Anchor-ViT, a novel Vision Transformer architecture for distracted driving detection. Anchor-ViT enhances the standard Vision Transformer framework [7] by integrating**learnable spatial anchors**and**Soft Radial Attention (SRA)**, drawing inspiration from anchor-based methods in computer vision [38], [39]. This innovative approach guides the model's attention towards driver-centric image regions – hands, face, interaction objects – using these anchors and SRA to capture subtle distraction cues, while leveraging the ViT backbone to maintain global scene context, ultimately improving detection accuracy.

A. *Vision Transformer Foundation*

Several key components of the ViT architecture [7] provide the essential basis for image processing and contextual understanding within Anchor-ViT. These foundational elements include patch embedding, where input images are divided into patches, linearly projected, and subsequently organized into an initial token sequence. Positional encoding is also incorporated, utilizing learned positional embeddings that are added to the patch embeddings to encode spatial information, which is crucial for maintaining spatial relationships within the image. Furthermore,

Anchor-ViT leverages a global context mechanism through a dedicated class token, denoted as [CLS]. This [CLS] token is designed to aggregate information globally across the image for classification purposes and, importantly, its attention mechanism remains unmodified by Anchor-ViT's innovations, ensuring the preservation of global context. Finally, the architecture includes Transformer encoder blocks, which are standard components comprising Multi-Head Self-Attention (MSA) and MLP modules and serve as the core feature extraction component, building upon established attention mechanisms [18], [19]. Anchor-ViT innovates upon this foundation by enhancing these Transformer encoder blocks with the Soft Radial Attention mechanism, which is detailed in subsequent sections. Collectively, these ViT components provide the necessary framework for image processing and global context integration, upon which Anchor-ViT's spatially guided attention mechanism is built.

B. *Learnable Spatial Anchors*

Anchor-ViT introduces a set of k learnable spatial anchors, denoted as $A = \{A_1, A_2, \ldots, A_k\}$, where each anchor $A_i = (A_{i,x}, A_{i,y}) \in \mathbb{R}^2$ represents a location in the patch-coordinate space of the input image. These anchors are not predefined but are directly optimized via gradient descent during training, effectively becoming adaptive focal points within the image. To balance computational efficiency with focused attention, a relatively small number of anchors is typically employed, with a practical range of $k = 3$ to 5. The initialization of these anchors is performed using one of two primary strategies. The first strategy, uniform distribution, involves distributing anchors somewhat evenly across the patch grid. To prevent symmetry and encourage independent learning among anchors, a small amount of random jitter is introduced to their initial positions. This uniform initialization strategy is intended to facilitate an initial exploration of the entire spatial domain of the input images. The second initialization strategy is strategic placement, where anchors are intentionally initialized near regions that are considered to be particularly relevant for distracted driving detection. For instance, anchors may be strategically placed near the bottom-center of the image, an area often associated with the presence of hands. Similar to the uniform distribution method, random jitter is also incorporated in this strategic initialization approach. It is crucial to note that the anchor positions are learnable parameters within the Anchor-ViT model and are updated across training epochs through standard backpropagation. Importantly, during a single forward pass of the network, these anchor positions remain static; there is no iterative refinement or per-layer adjustment of anchor positions during forward propagation. This design choice is made to maintain computational efficiency and architectural simplicity while still allowing the anchors to dynamically adapt to spatially relevant locations throughout the training process. As training progresses, the gradient descent optimization process guides the anchors to converge to positions that are most effective in maximizing the model's performance on the distracted driving detection task.

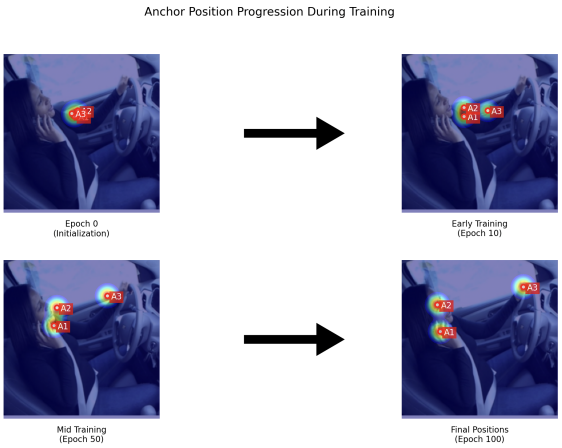

Fig. 1. Visualization of Anchor Position Progression During Training. The red dots with labels A1, A2, and A3 represent the learnable spatial anchors. The heatmap background indicates the radial weight distribution around the anchors. The anchors dynamically adjust their positions over epochs to focus on driver-relevant regions, starting from a uniform initialization to converging on areas like the face and hands.

C. *Soft Radial Attention (SRA)*

The Soft Radial Attention (SRA) mechanism, inspired by recent advancements in attention modeling [36], constitutes the core computational innovation of Anchor-ViT, functioning to leverage the learned anchors for modulating the self-attention process. For each token j, which represents a patch of the input image, and for each anchor i within the learned anchor set, the SRA mechanism calculates a radial weight, denoted as $w_{i,j}$. This calculation is based on a Gaussian kernel, formulated as

$$w_{i,j} = \exp\left(-\frac{\|(x_j, y_j) - A_i\|^2}{2\sigma^2}\right) \tag{1}$$

where (x_j, y_j) are the patch coordinates of token j, and $A_i = (A_{i,x}, A_{i,y})$ is the position of anchor i. The kernel bandwidth, σ, serves as a global parameter, which can be either fixed or learnable, and controls the spatial spread of each anchor's influence. A smaller value of σ leads to a more localized attention focus around each anchor. Following the computation of radial weights, these values are normalized to derive anchor-token distributions, denoted as $a_{i,j}$, according to the formula

$$a_{i,j} = \frac{w_{i,j}}{\sum_{i'=1}^{k} w_{i',j} + \varepsilon} \tag{2}$$

This normalization step ensures that for each token j, the sum of weights across all anchors approximates to unity, representing a soft, probabilistic assignment of each token to the anchor set. Within the Multi-Head Self-Attention (MSA) module, the SRA mechanism selectively modifies the attention logits, $A_{j,k}$, specifically for local-to-local token pairs, excluding interactions involving the CLS token to preserve global context. The adjustment is additive, expressed as

$$A_{j,k} \leftarrow A_{j,k} + \alpha \sum_{i=1}^{k} a_{i,j} \log(w_{i,k} + \varepsilon) \tag{3}$$

where the term $\sum_{i=1}^{k} a_{i,j} \log(w_{i,k} + \varepsilon)$ computes a weighted sum of log-radial weights. The scaling factor α, a hyperparameter that may be learnable, and the logarithmic function together control the magnitude and non-linearity of the radial influence on attention. This process effectively boosts attention between token pairs (j, k) that are spatially close to the same anchor(s) while suppressing attention between pairs distant from all anchors. Crucially, the attention associated with the CLS token remains unmodified, allowing it to maintain a global, anchor-agnostic view of the input, complementary to the localized, anchor-guided attention.

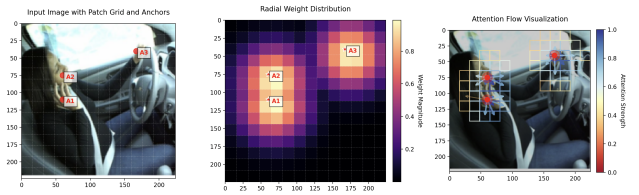

Fig. 2. Visualization of Soft Radial Attention (SRA) Mechanism. The figure comprises three parts: (Left) Input image with patch grid and anchor positions. (Middle) Radial weight distribution, showing higher weights (lighter colors) closer to the anchors. (Right) Attention flow visualization, illustrating how attention is concentrated around the anchors due to the SRA mechanism.

D. Loss Function

Anchor-ViT is trained end-to-end by minimizing a composite loss function designed to balance accurate classification and effective anchor learning. The total loss function, denoted as L, is a weighted combination of categorical cross-entropy loss (L_{CE}) and anchor regularization loss (L_{anchor}), such that $L = L_{CE} + L_{anchor}$. The Cross-Entropy Loss (L_{CE}) drives accurate distracted driving classification by quantifying the difference between predicted and ground-truth distraction class distributions. Complementing this, the Anchor Regularization Loss (L_{anchor}) ensures learned anchors are spatially diverse and stable, and is composed of the Repulsion Term (L_{repel}) and the Bounding Term (L_{bound}), where $L_{anchor} = L_{repel} + L_{bound}$. The Repulsion Term ($L_{repel}$), formulated as

$$L_{repel} = \lambda_{repel} \sum_{1 \le i < j \le k} \frac{1}{\|A_i - A_j\|^2 + \varepsilon} \tag{4}$$

encourages spatial diversity among anchors, preventing them from clustering redundantly and ensuring coverage of different relevant regions in the input image. The Bounding Term(L_{bound}), formulated as

$$L_{bound} = \lambda_{bound} \sum_{i=1}^{k} \max(0, \|A_i - A_i^{(init)}\|^2 - \delta) \tag{5}$$

promotes anchor stability during training, mitigating erratic anchor movement and ensuring they remain focused on learned, relevant spatial locations. This composite loss function guides Anchor-ViT training for both accurate classification and well-behaved anchor learning, leading to a more robust and interpretable model.

IV. EXPERIMENTS AND RESULTS

A. Datasets

Both the State Farm dataset [40] and 100-Driver dataset [41] were preprocessed by resizing images to 224×224 pixels. To enhance model robustness and generalization, standard data augmentation techniques were applied during training such as random resized cropping to 224×224, random horizontal flipping, and photometric augmentations including adjustments to brightness, contrast, and saturation.

Two distinct datasets were used:

1) **State Farm Distracted Driver Detection Dataset:** This widely-used benchmark [40] provides a rich set of driver images categorized into 10 classes representing various in-vehicle activities. These classes include safe driving and a range of driver distractions such as texting, talking on the phone, and operating the radio.
2) **100-Driver Distraction Dataset:** The 100-Driver dataset [41], originally comprising 22 fine-grained classes, offers a broader spectrum of real-world driving scenarios. To ensure compatibility with our 10-class output and allow direct comparisons with State Farm results, the dataset's original 22 classes were mapped to the 10 broader categories defined in the State Farm dataset, preserving the dataset's diversity while standardizing the classification task.

B. Implementation Details

Both models were implemented in PyTorch (v2.5) and trained on a single NVIDIA RTX 3090 GPU. We used SGD (momentum 0.9, weight decay 5e-5) with an initial learning rate of 0.001—scheduled via cosine annealing over 100 epochs, and a batch size of 64. The primary loss was categorical cross-entropy; for Anchor-ViT, an additional anchor regularization loss was applied ($\lambda_{repel} = 0.01$, $\lambda_{bound} = 0.01$, $\delta = 100$). Both models shared core hyperparameters: patch size 16, embedding dimension 384, 6 Transformer layers, and 6 attention heads per layer. Anchor-ViT specifically used 3 learnable spatial anchors, initialized uniformly over the patch grid with slight jitter (std. 1 patch unit), and a Soft Radial Attention mechanism configured with a fixed kernel bandwidth $\sigma = 0.5$ and weighting factor $\alpha = 1.0$.

C. Performance Results

TABLE I
SOTA COMPARISON ON STATE FARM AND 100-DRIVER DATASETS

Model	State-Farm Acc/F1 (%)	100-Driver Acc/F1 (%)
MobileVGG [8]	87.6 / 86.4	74.9 / 73.8
Swin-T [26]	88.3 / 89.1	79.4 / 80.2
FPT [27]	89.8 / 90.2	80.5 / 80.4
Li et al. hybrid [13]	90.1 / 90.3	81.6 / 82.1
Anchor-ViT (ours)	**92.3 / 92.4**	**83.4 / 83.7**

A broader comparison with state-of-the-art (SOTA) methods is in Table I. Over three seeds, Anchor-ViT yields

+2.2 $\pm$ 0.3 pp accuracy and +2.1 $\pm$ 0.4 pp F1 vs. the best prior method on State Farm, and +1.8 $\pm$ 0.2 pp accuracy on 100-Driver.

TABLE II
VALIDATION PERFORMANCE COMPARISON OF ANCHOR-VIT VS. BASELINE VIT ON THE STATE FARM AND 100-DRIVER DATASETS (*Metrics: Accuracy, Precision, Recall, and F1-Score*)

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Baseline ViT	State Farm	89.5	90.1	89.2	89.6
Anchor-ViT	State Farm	**92.3** ($\pm$0.3)	**92.8**	**92.1**	**92.4**
Baseline ViT	100-Driver	78.2	79.0	77.6	78.1
Anchor-ViT	100-Driver	**83.4** ($\pm$0.3)	**84.2**	**82.7**	**83.3**

As demonstrated in Table II, the Anchor-ViT model consistently outperforms the baseline ViT model. On the State Farm dataset, Anchor-ViT achieves a validation accuracy of **92.3%** (+2.8 pp vs. baseline). On the 100-Driver dataset, Anchor-ViT attains an accuracy of **83.4%** (+5.2 pp vs. baseline). The $\pm$0.3 pp accuracy variance (see Section IV-D) is noted.

D. Robustness to Anchor Placement

Anchor-ViT's robustness to anchor placement was evaluated. If anchors converge to poor locations, the learned SRA scaling factor $\alpha \to 0$, causing the network to revert to vanilla ViT; thus, accuracy should not drop below baseline. This was verified by training 15 configurations varying: five random seeds, two initialization schemes (uniform, face/hand priors), $\sigma \in \{0.4, 0.5, 0.6\}$, and $\lambda_{repel}, \lambda_{bound} \in \{0.005, 0.01, 0.02\}$. Across all runs, validation accuracy fluctuated by only $\pm$0.3 pp and was always above the ViT baseline. Anchor coordinates stabilized within the first five epochs (visualized in Fig. 1). This $\pm$0.3 pp variance is noted in Table II.

E. Ablation Studies

TABLE III
EFFECTIVENESS OF ANCHOR TYPE ON MODEL ACCURACY AND SPATIAL DIVERSITY

Model Variant	Accuracy (%)	Anchor Spatial Diversity (Avg. Euclidean Distance)
Baseline ViT (No Anchors)	78.2	N/A
Anchor-ViT (Fixed Anchors)	80.5	Fixed ($\approx$2.0)
Anchor-ViT (Learnable Anchors)	**83.4**	**2.6**

In Table III, the learnable anchors, which achieve an average spatial diversity of **2.6** patch units, yield the highest accuracy (**83.4%**) compared to the 78.2% and 80.5%. This suggests that learnable anchors enable the model to better capture spatially relevant features, contributing to improved performance.

TABLE IV
EFFECTIVENESS OF ANCHOR REGULARIZATION ON MODEL ACCURACY AND BOUNDING PENALTY

Model Variant	Accuracy (%)	Anchor Bounding Penalty (Avg. Value)
Anchor-ViT (No Regularization)	81.7	0.0
Full Anchor-ViT (With Regularization)	**83.4**	**0.27**

In Table IV, When the regularization term is applied, a small average bounding penalty (0.27) is incurred, but accuracy improves to **83.4%** from 81.7% indicating that constraining the anchors during training contributes to better generalization.

TABLE V
EFFECTIVENESS OF SOFT RADIAL ATTENTION (SRA) ON MODEL ACCURACY AND ATTENTION MAP ENTROPY

Model Variant	Accuracy (%)	Attention Map Entropy (Average)
Anchor-ViT (No SRA)	82.0	4.35
Full Anchor-ViT (With SRA)	**83.4**	**3.72**

In Table V, without SRA, attention entropy is 4.35 with 82.0% accuracy. With SRA, entropy drops to 3.72 and accuracy rises to **83.4%**, showing SRA sharpens focus on key spatial regions, improving distracted driving detection.

V. CONCLUSION

This investigation into the Anchor-ViT design—an extension of the Vision Transformer (ViT) framework [7]—has demonstrated significant improvements in detecting subtle, spatially specific signs of distraction. By integrating learnable spatial anchors and Soft Radial Attention (SRA), inspired by recent advances in attention mechanisms [36], Anchor-ViT outperformed standard ViT models, achieving a 2.8% improvement on the State Farm dataset [40] and a notable 5.2% boost on the more challenging 100-Driver dataset [41]. Although the core ViT architecture maintains comparable general image processing capabilities, the addition of anchor learning and radial attention adjustment enhances the focus on relevant regions—a crucial factor in detecting distracted driving. Future work will explore lightweight versions of Anchor-ViT, potentially drawing inspiration from efficient ViT architectures like AttnZero [20], through techniques such as model distillation and novel architectural approaches, aiming to reduce computational demands and enable real-time deployment in automotive contexts. Additionally, given that both standard ViT and Anchor-ViT treat images as individual snapshots, future research will investigate integrating temporal context using video-centric transformer techniques, similar to DRVMon-VM [31], or other innovative adaptations to better capture distraction behaviors as they develop over time.

REFERENCES

[1] National Center for Statistics and Analysis, "State traffic data: 2021 data (traffic safety facts. report no. dot hs 813 509)," September 2023, national Highway Traffic Safety Administration.

[2] H. V. Koay, J. H. Chuah, C.-O. Chow, and Y.-L. Chang, "Detecting and recognizing driver distraction through various data modality using machine learning: A review, recent advances, simplified framework and open challenges (2014–2021)," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105309, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197622003517

[3] H. Gao and Y. Liu, "Improving real-time driver distraction detection via constrained attention mechanism," *Engineering Applications of Artificial Intelligence*, vol. 128, p. 107408, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197623015920

[4] Q. Cheng, H. Li, Y. Yang, J. Ling, and X. Huang, "Towards efficient risky driving detection: A benchmark and a semi-supervised model," *Sensors*, vol. 24, no. 5, p. 1386, Mar. 2024. [Online]. Available: https://doi.org/10.3390/s24051386

[5] D. M. Pisharody, B. P. Chacko, and K. M. Basheer, "Driver distraction detection using machine learning techniques," *Materials Today: Proceedings*, vol. 58, no. Part 1, pp. 251–255, 2022, accessed from https://doi.org/10.1016/j.matpr.2022.02.108.

[6] N. K. Vaegae, K. K. Pulluri, K. Bagadi, and O. O. Oyerinde, "Design of an efficient distracted driver detection system: Deep learning approaches," *IEEE Access*, vol. 10, pp. 116 087–116 097, 2022.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[8] B. Baheti, S. Talbar, and S. Gajre, "Towards computationally efficient and realtime distracted driver detection with mobilevgg network," *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 4, pp. 565–574, 2020.

[9] F. Sajid, A. R. Javed, A. Basharat, N. Kryvinska, A. Afzal, and M. Rizwan, "An efficient deep learning framework for distracted driver detection," *IEEE Access*, vol. 9, pp. 169 270–169 280, 2021.

[10] B. Ma, Z. Fu, S. Rakheja, D. Zhao, W. He, W. Ming, and Z. Zhang, "Distracted driving behavior and driver's emotion detection based on improved yolov8 with attention mechanism," *IEEE Access*, vol. 12, pp. 37 983–37 994, 2024.

[11] Z. Wang and L. Yao, "Recongnition of distracted driving behavior based on improved bi-lstm model and attention mechanism," *IEEE Access*, vol. 12, pp. 67 711–67 725, 2024.

[12] R. Wang, L. Huang, and C. Wang, "Fast detection of handheld phone-distracted driving by sensing the driver's hand-grip," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 8, pp. 11 136–11 149, 2024.

[13] Z. Li, X. Zhao, F. Wu, D. Chen, and C. Wang, "A lightweight and efficient distracted driver detection model fusing convolutional neural network and vision transformer," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 19 962–19 978, 2024.

[14] P. Li, Q. Mou, J. Hou, and Y. Tu, "Hybrid convolutional-transformer neural network for driver distraction detection," in *2023 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS)*, 2023, pp. 1–6.

[15] M. A. R. Mollah, A. Y. Srizon, and M. Ahmed, "An ensemble approach for identification of distracted driver by implementing transfer learned deep cnn architectures," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, 2022, pp. 938–942.

[16] Y. Liu, S. Du, Q. Guo, Z. Zhao, Z. Tian, and N. Zheng, "Structure consistent unsupervised domain adaptation for driver behavior recognition," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2023, pp. 1038–1043.

[17] H. Yang, H. Liu, Z. Hu, A.-T. Nguyen, T.-M. Guerra, and C. Lv, "Quantitative identification of driver distraction: A weakly supervised contrastive learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 2034–2045, 2024.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016. [Online]. Available: https://arxiv.org/abs/1409.0473

[20] L. Li *et al.*, "Attnzero: Efficient attention discovery for vision transformers," in *Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol 15063*. Springer, Cham, 2025.

[21] H. Chen, H. Liu, X. Feng, and H. Chen, "Distracted driving recognition using vision transformer for human-machine co-driving," in *2021 5th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, 2021, pp. 1–7.

[22] Z. Fang, J. Chen, J. Wang, Z. Wang, N. Liu, and G. Yin, "Driver distraction behavior detection using a vision transformer model based on transfer learning strategy," in *2022 6th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, 2022, pp. 1–6.

[23] H. V. Koay, J. H. Chuah, and C.-O. Chow, "Convolutional neural network or vision transformer? benchmarking various machine learning models for distracted driver detection," in *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, 2021, pp. 417–422.

[24] Y. Li, L. Wang, W. Mi, H. Xu, J. Hu, and H. Li, "Distracted driving detection by combining vit and cnn," in *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2022, pp. 908–913.

[25] H. V. Koay, J. H. Chuah, and C.-O. Chow, "Shifted-window hierarchical vision transformer for distracted driver detection," in *2021 IEEE Region 10 Symposium (TENSYMP)*, 2021, pp. 1–7.

[26] C. Song, Q. Song, and F. Cao, "Driver distraction detection based on improved swin transformer," in *2024 43rd Chinese Control Conference (CCC)*, 2024, pp. 8032–8037.

[27] H. Wang, J. Chen, Z. Huang, B. Li, J. Lv, J. Xi, B. Wu, J. Zhang, and Z. Wu, "Fpt: Fine-grained detection of driver distraction based on the feature pyramid vision transformer," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 1594–1608, 2023.

[28] Y. Wang, Z. Li, G. Guan, Y. Sun, C. Wang, H. R. Tohidypour, P. Nasiopoulos, and V. C. Leung, "Weighted multi-task vision transformer for distraction and emotion detection in driving safety," in *2024 International Conference on Computing, Networking and Communications (ICNC)*, 2024, pp. 152–156.

[29] Y. Ma, L. Yuan, A. Abdelraouf, K. Han, R. Gupta, Z. Li, and Z. Wang, "M2dar: Multi-view multi-scale driver action recognition with vision transformer," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 5287–5294.

[30] G. Ding, W. Han, C. Wang, M. Cui, L. Zhou, D. Pan, J. Wang, J. Zhang, and Z. Chen, "A coarse-to-fine boundary localization method for naturalistic driving action recognition," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 3233–3240.

[31] Y. Ma, R. Du, A. Abdelraouf, K. Han, R. Gupta, and Z. Wang, "Driver digital twin for online recognition of distracted driving behaviors," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 2, pp. 3168–3180, 2024.

[32] T. Zhang, Q. Wang, X. Dong, W. Yu, H. Sun, X. Zhou, A. Zhen, S. Cui, D. Wu, and Z. He, "Augmented self-mask attention transformer for naturalistic driving action recognition," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024, pp. 7108–7114.

[33] C. Duan, J. Liao, N. Ding, and L. Cao, "Enabling robust distracted driving performance across datasets with clip," in *2023 2nd International Conference on Sensing, Measurement, Communication and Internet of Things Technologies (SMC-IoT)*, 2023, pp. 112–116.

[34] A. Sonth, A. Sarkar, H. Bhagat, and L. Abbott, "Explainable driver activity recognition using video transformer in highly automated vehicle," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, 2023, pp. 1–8.

[35] R. Rezaei, M. Jalili Sabet, J. Gu, D. Rueckert, P. Torr, and A. Khakzar, "Learning visual prompts for guiding the attention of vision transformers," 2024. [Online]. Available: https://arxiv.org/abs/2406.03303

[36] T. Nguyen, C. D. Heldermon, and C. Toler-Franklin, "Unified local and global attention interaction modeling for vision transformers," 2024. [Online]. Available: https://arxiv.org/abs/2412.18778

[37] B. Li, J. Chen, Z. Huang, H. Wang, J. Lv, J. Xi, J. Zhang, and Z. Wu, "A new unsupervised deep learning algorithm for fine-grained detection of driver distraction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19 272–19 284, 2022.

[38] X. Zhang, F. Wan, C. Liu, X. Ji, and Q. Ye, "Learning to match anchors for visual object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3096–3109, 2022.

[39] B. Jiang, S. Luo, X. Wang, C. Li, and J. Tang, "Amatformer: Efficient feature matching via anchor matching transformer," *IEEE Transactions on Multimedia*, vol. 26, pp. 1504–1515, 2024.

[40] "State farm distracted driver detection," [Online], 2016, accessed 2-January-2025. Available: https://www.kaggle.com/c/state-farm-distracted-driver-detection/overview.

[41] J. Wang, W. Li, F. Li, J. Zhang, Z. Wu, Z. Zhong, and N. Sebe, "100-driver: A large-scale, diverse dataset for distracted driver classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 7061–7072, 2023.