Integration of Large Vision Models in Driver Monitoring Systems: Compressing and Distilling for Real-Time Automotive Applications

Georgios Markos Chatziloizos
Renault Group - Ampere
University Cote d'Azur / CNRS-I3S
Nice, France
georgios.chatziloizos@renault.com

Andrea Ancora
Renault Group - Ampere
Nice, France
andrea.ancora@renault.com

Andrew I. Comport
University Cote d'Azur / CNRS-I3S
Nice, France
Andrew.Comport@cnrs.fr

Christian Barat
University Cote d'Azur / CNRS-I3S
Nice, France
barat@i3s.unice.fr

Abstract

This study focuses on optimizing neural network architectures for real-time detection of driver facial bounding boxes. Initially, we trained the Florence-2 model, which demonstrated high accuracy but proved too large for real-time applications. To address this, we employed model distillation, using Florence-2 as a teacher to train a more compact DINOv2 model. Our aim was to maintain high detection accuracy while minimizing memory usage and inference time, making the solution viable for real-time implementation on GPU and NPU devices. We present a comparative analysis of model performance in terms of IoU scores, memory consumption and inference times.

1 Introduction

Real-time driver monitoring systems play an essential role in ensuring the safety and comfort of both drivers and passengers. These systems are integrated into various applications, such as driver behavior analysis [11, 2], fatigue detection and emotion recognition [8, 13], which all depend on accurate and low-latency detection of the driver's face. Detecting the driver's face allows the system to track eye movements [1, 3], monitor head pose [5, 10] and identify expressions or behaviors that may indicate distraction or drowsiness. Given the critical nature of these applications, achieving a balance between detection accuracy and computational efficiency is a significant challenge.

Accurate face detection often relies on complex deep learning models that are computationally intensive. However, deploying such models in real-time environments, especially on edge devices with limited computational resources, introduces several bottlenecks. Edge devices, such as smartphones, in-car processors and embedded systems, often lack the memory and processing power available on larger platforms like servers or high-end GPUs. These constraints demand models that are both lightweight and fast while maintaining high accuracy in detecting driver facial regions in a variety of conditions, including different lighting, facial orientations and occlusions.

The Florence-2 model [12] (MIT License) is a state-of-the-art vision model known for its strong performance on various visual tasks, including face detection. With its deep architecture and extensive pretraining, Florence-2 achieves high accuracy, making it a strong candidate for driver face detection.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

However, its large memory footprint (over 1 GB) and relatively slow inference time (around 90 ms on a GPU) render it impractical for real-time deployment on resource-constrained devices. In real-time applications, even small increases in latency can significantly impact the user experience or the system's reliability, particularly in safety-critical scenarios such as driver monitoring.

To address these limitations, we explored the use of model compression techniques, specifically knowledge distillation[4]. Knowledge distillation is a popular method for transferring knowledge from a large, complex "teacher" model (in this case, Florence-2) to a smaller, more efficient "student" model. The student model, while smaller and faster, is trained to mimic the output of the teacher model, allowing it to achieve similar levels of accuracy. By using Florence-2 as the teacher, we trained a distilled version of the DINOv2 model [9] (Apache License 2.0). DINOv2 is known for its compact architecture (with only 22 million parameters) and efficiency, making it a suitable candidate for edge deployment.

The goal of this work was to reduce the memory footprint and inference time of the face detection model while maintaining high detection accuracy, particularly in terms of the Intersection over Union (IoU) score for bounding box detection. By distilling the Florence-2 model into DINOv2, we were able to produce a smaller, faster model that still performs well in real-time driver face detection tasks.

In this paper, we provide a detailed comparison of the performance of the original Florence-2 model, a quantized version of Florence-2 and the DINOv2 model. We evaluate these models in terms of their accuracy (measured by IoU scores), memory usage and inference speed on both GPU and NPU hardware. Our results demonstrate the effectiveness of model distillation in achieving real-time performance on edge devices without significantly compromising detection accuracy.

2 Process

Integrating a large vision model, such as Florence-2, into an in-car system is essential because of its ability to deliver high precision and comprehensive monitoring of the driver's face and behavior. Larger models are typically trained on extensive datasets and possess deep architectures, enabling them to capture fine details and variations in driver states, such as subtle facial expressions, eye movements and head pose, even under challenging conditions like low light or partial occlusions. This precision is crucial in safety-critical applications, where accurate detection of fatigue, distraction, or emotional states can directly impact driver safety and response times.

However, the size and complexity of large models make them impractical for real-time deployment on resource-limited automotive hardware. Real-time systems in cars, such as NPUs or edge GPUs, have stringent memory and processing constraints. A large model, with its substantial memory footprint and slower inference speed, would introduce latency and consume excessive resources, hindering the system's ability to respond quickly.

In our case, the optimization process for driver face detection involved several critical stages. We began by utilizing a proprietary, non-public dataset, which consisted from synthetic infrared images and contained labeled bounding boxes around drivers' faces. This versatile dataset was designed to simulate diverse real-world conditions, including variations in lighting, angles and driver positions. Subsequently, the Florence-2 model was fine-tuned using this dataset to maximize its detection accuracy. However, despite the model's robust performance, its substantial size and slow inference time made it unsuitable for real-time use. In an effort to address this, we applied 4-bit quantization to reduce the model's memory footprint. Unfortunately, this further slowed the inference time, making the model less viable for real-time deployment.

To address the slow inference time issue, we employed knowledge distillation, which is a process that transfers knowledge from a large, complex model, known as the teacher, to a student model. As shown in Figure 1, the teacher model is Florence-2 and the student model is DINOv2. The process begins with an input image of size 1020 x 1366 pixels. This image is fed into the teacher model, Florence-2, which processes the image and generates teacher predictions.

Simultaneously, the ground truth data, which represents the correct or desired output, is used as a reference. The teacher predictions and the ground truth are compared to calculate the distillation loss, which measures the discrepancy between the teacher's predictions and the ground truth.

The input image is also resized to 224 x 224 pixels and fed into the student model, DINOv2. This model processes the image and generates student predictions. These student predictions are compared with the ground truth to further refine the student model. Furthermore, the Adam optimizer [6] is utilized to tune the model's parameters with starting learning rate $2 * 10^{-4}$ and using cosine annealing [7] to modify it in the last epochs.

Despite being smaller and less complex, the student model captures the essential knowledge encoded by the teacher. By optimizing for efficiency, the student model becomes much lighter in terms of memory usage and computational demand, making it suitable for real-time applications like driver face detection on edge devices. Although the student model may sacrifice some accuracy compared to the teacher, it gains speed and scalability, achieving a balance between performance and efficiency.

The distillation loss is used to guide the training of the student model, DINOv2, by transferring knowledge from the teacher model, Florence-2. This process aims to improve the performance of the student model by leveraging the knowledge of the pre-trained teacher model.

The distillation loss L_{distill} is calculated using the SmoothL1 loss between the student model's output p_s and the teacher model's output p_t:

$$L_{\text{distill}} = \text{SmoothL1}(p_s, p_t)$$

The ground truth loss L_{gt} is similarly calculated using the SmoothL1 loss between the student model's output p_s and the ground-truth labels y:

$$L_{\text{gt}} = \text{SmoothL1}(p_s, y)$$

Also, the IoU loss is calculated as:

$$L_{\text{iou}} = 1 - \text{IoU}(p_s, y)$$

The total loss is a combination of the distillation loss, the ground-truth loss and the IoU loss, weighted by the parameter α and β for the IoU loss:

$$L_{\text{total}} = \alpha \cdot L_{\text{distill}} + (1 - \alpha) \cdot L_{\text{gt}} + \beta \cdot L_{\text{iou}}$$

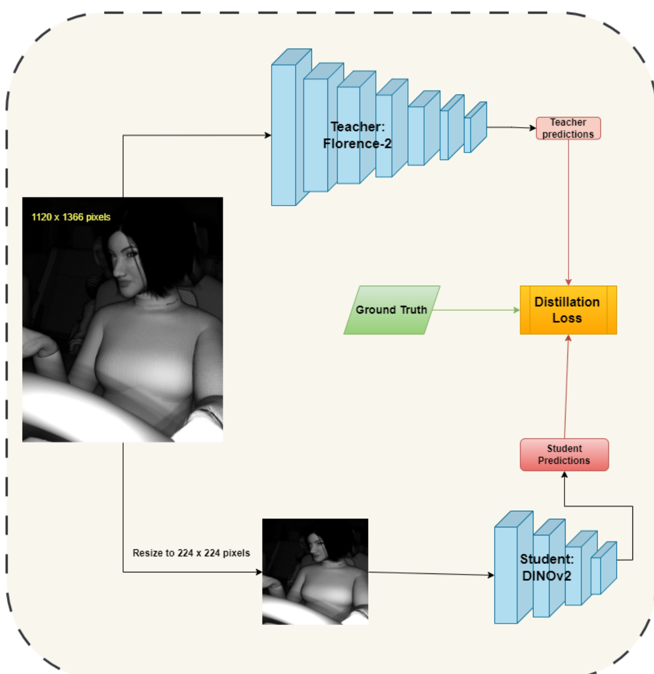

Figure 1: Knowledge distillation from Florence-2 to DINOv2

3 Architectures and Results

We evaluated the performance of three models, Florence-2, Florence-2 4-bit Quantized and DINOv2 Distilled, on the task of detecting drivers' faces which were trained on A6000 Nvidia GPUs. The results, shown in Table 1, illustrate the trade-offs between model size, IoU score and inference time on A6000 Nvidia GPU and on the QC SA8255P NPU hardware which is specifically designed for automotive applications.

Table 1: IoU Scores and Time for Different Models

Model	IoU Score	Memory size (in MB)	GPU Time (in ms)	NPU Time (in ms)
Florence-2	**0.980**	1009.1	90.0	-
Florence-2 4-bit-Quantized	0.978	274.5	110.2	-
DINOv2	0.913	**84.7**	**7.7**	11.5

As shown in Table 1, Florence-2 achieved the highest IoU score of 0.980 but at the cost of significant memory usage and slower inference time. Quantizing Florence-2 to 4-bit precision reduced its size but with a minimal decrease in IoU. The DINOv2 model, while slightly lower in accuracy with an IoU of 0.913, drastically reduced memory usage to 84.7 MB and significantly improved inference speed. Notably, when deployed on an automotive-grade NPU, the DINOv2 model achieved a real-time performance of 87 frames per second (FPS), making it highly suitable for in-car systems that require rapid and reliable face detection. We attribute the lower IoU scores not solely to architectural differences but primarily to the reduced image resolution. While distillation at the native resolution for DINOv2 is impractical, we believe the model's performance remains sufficiently robust at the current resolution.

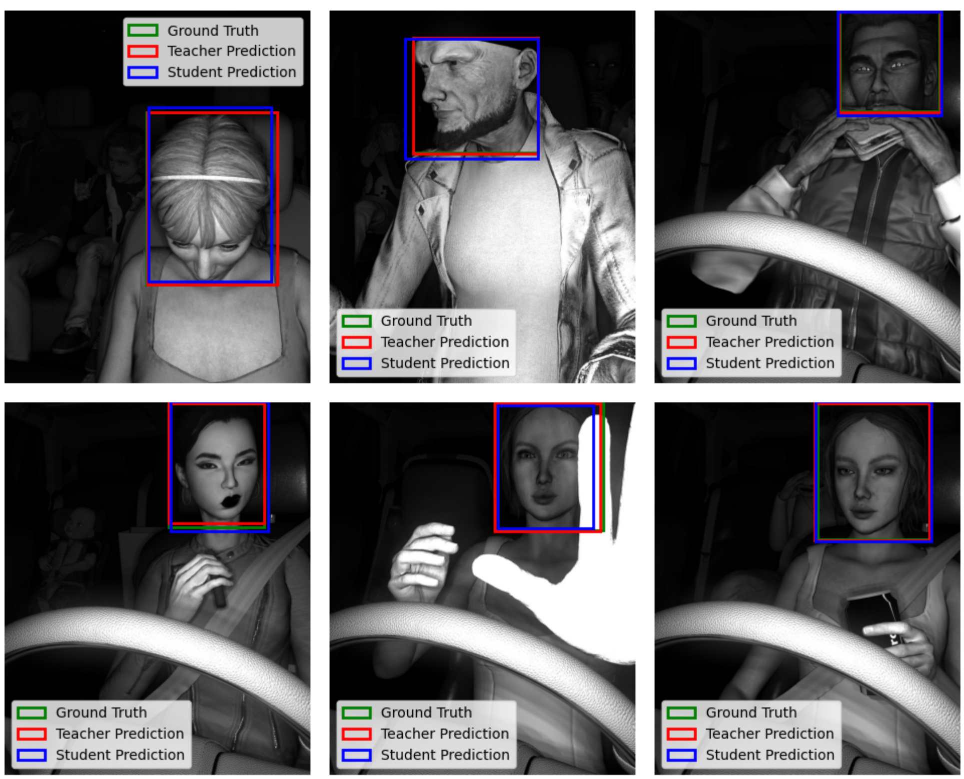

Figure 2: Example Bounding boxes of Driver's face for Student, Teacher and Ground Truth

The results displayed in Figure 2 showcase the bounding boxes for driver face detection produced by the student model, teacher model and ground truth labels. The green boxes represent the ground truth, red boxes represent the teacher's predictions and blue boxes represent the student's predictions.

Across multiple images, the student model's bounding boxes (in blue) closely align with the teacher's predictions (in red), suggesting effective knowledge transfer through model distillation. Despite minor discrepancies in certain frames, such as variations in face angles and lighting conditions, the student model demonstrates strong performance, closely approximating the accuracy of the teacher model while being computationally more efficient.

4 Conclusions

This study demonstrates the trade-offs between accuracy and efficiency in real-time driver face detection models. Although Florence-2 achieves superior accuracy, its large size and slow inference make it unsuitable for real-time applications on edge devices. Using knowledge distillation, we successfully trained the DINOv2 model, which, despite a slight drop in accuracy, achieved remarkable reductions in memory usage and inference time. This makes DINOv2 a viable solution for real-time driver face detection, particularly in resource-constrained environments like NPUs.

Future work will explore further optimizations, including quantization and hardware-specific tuning, to improve both accuracy and efficiency even further. Also, we plan to focus on enhancing the system's ability to track driver emotions, behaviors and patterns more effectively in real-time. By improving the model's ability to recognize subtle facial expressions and gestures, we aim to provide deeper insights into the driver's emotional state and overall driving behavior. This could lead to improved safety interventions and a more comprehensive understanding of driver dynamics.

Acknowledgments and Disclosure of Funding

The authors would like to acknowledge the Association Nationale Recherche Technologie (ANRT) for CIFRE funding (2023/0751).

References

[1] Dario Babić, Helena Dijanić, Lea Jakob, Darko Babić, and Eduardo Garcia-Garzon. Driver eye movements in relation to unfamiliar traffic signs: An eye tracking study. *Applied ergonomics*, 89:103191, 2020.

[2] Georgios Markos Chatziloizos, Andrea Ancora, Andrew I. Comport, and Christian Barat. Low parameter neural networks for in-car distracted driver detection. In *Proceedings of the 2024 9th International Conference on Machine Learning Technologies*, ICMLT '24, page 204–208, New York, NY, USA, 2024. Association for Computing Machinery.

[3] Martin Eriksson and Nikolaos P Papanikotopoulos. Eye-tracking for detection of driver fatigue. In *Proceedings of Conference on Intelligent Transportation Systems*, pages 314–319. IEEE, 1997.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[5] Sumit Jha and Carlos Busso. Analyzing the relationship between head pose and gaze to model driver visual attention. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2157–2162. IEEE, 2016.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[7] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

[8] Rizwan Ali Naqvi, Muhammad Arsalan, Abdul Rehman, Ateeq Ur Rehman, Woong-Kee Loh, and Anand Paul. Deep learning-based drivers emotion classification system in time series data for remote applications. *Remote Sensing*, 12(3):587, 2020.

[9] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat,

Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[10] Anke Schwarz, Monica Haurilet, Manuel Martinez, and Rainer Stiefelhagen. Driveahead-a large-scale driver head pose dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10, 2017.

[11] Mohammad Shahverdy, Mahmood Fathy, Reza Berangi, and Mohammad Sabokrou. Driver behavior detection and classification using deep convolutional neural networks. *Expert Systems with Applications*, 149:113240, 2020.

[12] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*, 2023.

[13] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W Picard. Driver emotion recognition for intelligent vehicles: A survey. *ACM Computing Surveys (CSUR)*, 53(3):1–30, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The experiments and the figures show that we have successfully disstilled DINOv2 model and it can be used in real time applications for in-car monitoring

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: There is a performance loss when doing distillation particularly because of the smaller size of the student model, but also for resizing the images to a smaller scale.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes all the information in order to reproduce it are provided, the models, how to train them, hyperparameters etc.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code will be available soon. Unfortunately, the data are proprietary and at this moment cannot be made available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, these details were specified in section Process

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This work it is just a preliminary work, in future work error bars will be added.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We explained which GPUs and NPUs we used for training but also doing inference.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, this preliminary work explores ways to make autonomous and assistive driving safer. We used diverse synthetic data in order to train our models.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The models are trained on synthetic dataset consisted of characters of different ethnicities, ages etc.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Both DINOv2 and Florence-2 models's licences are stated (Apache License 2.0 and MIT License, respectively).

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, the code and documentation will be provided in the next stages.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The data are synthetic.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The data are synthetic.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.