# Video prediction with pruned transformer

Stanford CS229 Project

**Dongyeong Kim**
Department of Computer Science
Stanford University
dkim1232@stanford.edu

## Abstract

In the burgeoning domain of computer vision, advancements proliferate across a spectrum from rudimentary object detection and lane tracing to sophisticated image generation. A notable trend is the emphasis on ascertaining the present state of an image, a process that inherently introduces latency between observation and optimized decision-making. Within the ambit of this study, we delineate the development of a model poised for the prediction of future imagery based on antecedent video frames. This paradigm shift towards predictive analysis necessitates rapid computational processes, underpinning our selection of three distinct methodologies for video frame examination: Convolutional Neural Networks (CNN), an Innovative Transformer for Temporal Resolution (ITTR), and Sequential Transformers.

The CNN, a staple in the feature extraction phase, is harnessed for its proficiency in delineating salient features within images. In tandem, the Sequential Transformer is deployed for its alacrity in analyzing sequences of frames, a critical component in the temporal dimension of our model. Distinctively, the ITTR transformer, an evolution of the conventional image transformer, eschews the standard patch embedding technique. Instead, it adopts a selective approach towards the height and width of an image, focusing exclusively on elements deemed pertinent. This pruning mechanism ensures a concentrated update on salient image features, thereby facilitating a nuanced analysis of the image context with the transformer's architecture.

The synthesis of these methodologies within the CNN-ITTR-Sequential Transformer model enables a comprehensive analysis encompassing image features, contextual nuance, and sequential importance. The ensuing upscaling of results in our experimental forays underscores the model's efficacy in preempting future states of video frames, heralding a new vista in predictive computer vision technology.

## 1 Introduction

The impetus for this project emanates from the quest to mitigate potential hazards in autonomous vehicles through advanced imaging techniques. Conventional Anti-lock Braking Systems (ABS) equipped with distance sensors or cameras operate on immediate inputs and are fundamentally reactive. While the rapid responses facilitated by these sensors and ASIC devices are commendable, they only activate upon imminent danger, placing the vehicle and its occupants in precarious situations. This realization spurred the development of a model designed to foresee future events and contexts, thereby averting potential dangers before they evolve into emergencies.

To elaborate, the model harnesses the power of Convolutional Neural Networks (CNNs), Image Transformers (ITTR), and sequential transformers. The initial stage employs a CNN to process input images structured as [batch, number of frames, 3 (channels), height, width], utilizing 8 batches, 4 frames, and a resolution of 224x224 pixels, tailored to the computational limits of my setup. Through

a tripartite process of CNN downsampling and channel augmentation, the dimensions are refined to [8, 4, 256, 28, 28].

Subsequently, these frame images are analyzed by an ITTR transformer, necessitating a tensor reshaping to [32, 256, 28, 28] for input. Herein, a weight-applied image is merged with the initial convoluted image without pooling and subjected to further convolution to restore the original dimensions, resulting in an ITTR output of [32, 256, 28, 28].

This output undergoes temporal embedding to discern frame-to-frame correlations, resizing the tensor to [32, 2562828] for encoder attention. However, the enormity of encoding this size surpassed the computational bounds of my system, leading to dimensionality reduction via random projection, thus resizing the input to [32, 1000], with the output maintaining equivalent dimensions. The processed results are then fed through a transformer encoder and a linear transformer, reverting the dimensions to [32, 2562828].

To synthesize the ITTR and sequential transformer outcomes, the result is kept at [8, 4, 256, 28, 28]. For final prediction generation, the dimension is adjusted to [8, 1024, 28, 28] and subjected to an upsampling process, yielding an image of dimensions [8, 3, 28, 28]. This integrative multi-model approach, predicated on image generation, employs Generative Adversarial Network (GAN) loss to validate the generated image's ability to encapsulate features through sequential interconnections, offering a proactive solution to vehicular safety.

## 2 Related Work

This research draws inspiration from seminal works in the domain of image processing and neural networks, specifically referencing the Image Transformer (ITTR)(Zheng et al., 2022).” , "Attention Is All You Need"(Vaswani et al., 2023).”, and the integration of Convolutional Neural Networks (CNNs) and transformers for object detection(Vaswani et al., 2023).”. ITTR is lauded for its computational efficiency, employing a targeted approach to attention that prioritizes significant segments of the image over conventional pixel or patch-based attention mechanisms. However, ITTR's selective focus comes with the inherent drawback of potential data loss by neglecting less prominent image areas. Distinguishing from ITTR's application in singular image transformations, the proposed model innovates by incorporating sequential embedding to discern connections across frames, enhancing the comprehension of temporal dynamics.

The methodological foundation for temporal embedding is adapted from the transformative principles outlined in "Attention Is All You Need" and the amalgamation of CNNs and transformers for object detection, advocating for a nuanced understanding of sequential relationships. This approach is adept at capturing both immediate and extended temporal contexts, presenting a significant leap over traditional sequential models like RNNs or LSTMs. Despite its advantages, the approach is not without its challenges, notably the requisite for extensive datasets and computational resources. To circumvent these limitations, I employed dimensionality reduction through random projections, guided by the principles of the Johnson-Lindenstrauss Lemma and findings in "Experiments with Random Projection"(Dasgupta, 2013).”. This strategy enabled a reduction in data dimensionality from 2562828 to 1000, incurring an error boundary of approximately 13.8 percent compared to the original dataset, thereby maintaining a balance between computational feasibility and data integrity.

Furthermore, the integration of UNet principles, as derived from the ResUNet-a framework(Diakogiannis et al., 2020).”, significantly enhances the model's capability in image upsampling. This adaptation is particularly pertinent for synthesizing predictive frames that are closely aligned with the immediate predecessor, leveraging last-frame information through concatenation to refine the upsampling process. Nonetheless, the UNet architecture has its limitations, especially in scenarios characterized by abrupt frame changes or the introduction of unforeseen elements. Despite these constraints, the model is predicated on forecasting within the visible spectrum, adeptly generating plausible future scenarios.

This holistic model merges the strengths of ITTR's focused computational approach, the temporal depth of sequential transformers, and the spatial fidelity of UNet-based upsampling. By addressing the inherent limitations of each component through strategic modifications and integrations, the proposed model stands as a robust framework for anticipating future states, marking a significant stride towards proactive safety measures in autonomous vehicle navigation.

## 3 Dataset and Features

The dataset underpinning this study is derived from authentic driving scenarios sourced from YouTube videos and the renowned KITTI dataset, developed collaboratively by the Karlsruhe Institute of Technology and Toyota Technological Institute. Esteemed research works such as "DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving"(Chen et al., 2015).", "Monocular Depth Estimation Based on Deep Learning"(Zhao et al., 2020).", and "MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving"(Teichmann et al., 2018)." have previously leveraged the KITTI dataset for its comprehensive representation of driving conditions. Additionally, investigations documented in "Learning to Drive in a Day"(Kendall et al., 2018)." and The Oxford RobotCar Dataset have utilized extensive driving data from the Oxford RobotCar, spanning over a year. Similarly, "Fast and Furious: Real-Time End-to-End 3D Detection, Tracking, and Motion Forecasting with a Single Convolutional Net"(Luo et al., 2020)." incorporated real driving data into their research, highlighting the prevalent use of authentic driving scenarios in contemporary studies, including a diverse array of driving conditions found in YouTube videos from locales such as Los Angeles, Orange County, Austin, and Oregon.

Adapting this dataset for model training presented several challenges, primarily due to limitations in cloud computing resources. The sheer volume of data significantly extended training durations, while high-resolution imagery, though beneficial for detailed analysis, demanded substantial computational power and memory. To mitigate these constraints, the resolution was adjusted to 224x224 pixels, forming a dataset comprising 8000 training sets, with an additional 400 sets allocated for validation and testing, based on experimental determinations of the requisite volume for effective regularization.

The dataset composition includes four consecutive frames as inputs and a single frame as the label, corresponding to the subsequent 0.2-second interval. This configuration was inspired by a methodical extraction from YouTube driving footage and the KITTI dataset, selecting sequences of five frames at 0.2-second intervals. Such an approach, depicted in figure 1, utilizes continuous input frames over a 0.8-second span with the labeling image representing the immediate future frame. Optimized for cloud computing efficiencies, this methodology avoids the cumbersome process of individually downloading and uploading data by enabling the use of extensive 1-2 hour driving recordings to extract realistic dataset scenarios. However, despite these optimizations, the dataset's original form required refinement to ensure suitability for model training, including adjustments to represent various road conditions, albeit within the constraints of available climate and road condition variability.

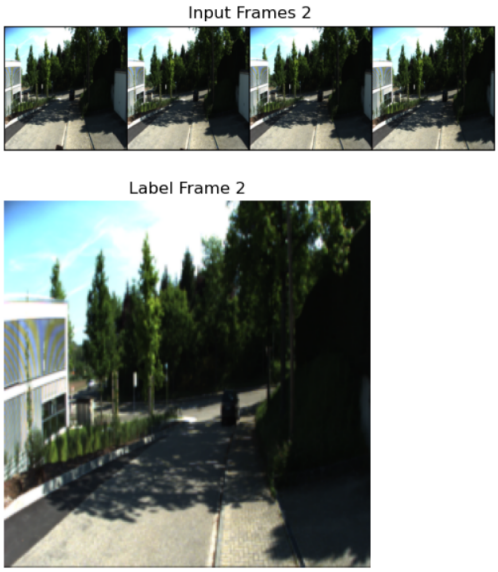


Figure 1: example 4 input and label from KITTI dataset


Figure 2: example 4 input and label from YouTube driving video

# 4 Methods

## 4.1 ITTR(Unpaired Image-to-Image Translation with Transformers)

The overall architecture is composed with CNN downsampling convolution, hybrid perception block. The CNN block can capture features and with overlapping patch embedding, it can provide a wide range of data, and context. And HPB block architecture can provide complex context based on the data from CNN convolution.

$$\text{Stem}(I), \quad X \in \mathbb{R}^{H/8 \times W/8 \times C} \tag{1}$$

The stem($I$) is the result of CNN data that in my model, the CNN is done with 3 stages of 2 stride downsampling. This makes the height and width decrease and increase the channel.

$$X = HPB_i(X), \quad i \in \{1, ..., 9\}, \quad X \in \mathbb{R}^{H/4 \times W/4 \times C} \tag{2}$$

And the result X is used as the input of the HPB block. The figure3 is the architecture of HPB block. In the DPSA, the ITTR calculates the contribution of tokens grouped by rows or columns. Therefore,

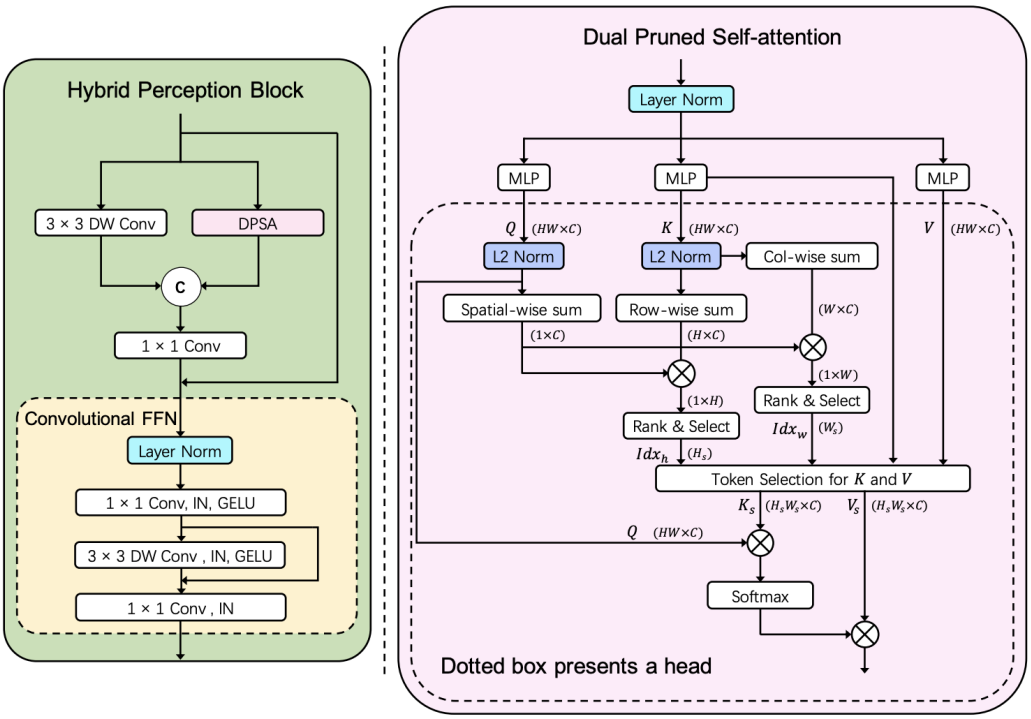


Figure 3: the overall architecture of ITTR, the architecture of hybrid perception block (HPB), dual pruned self-attention (DPSA). "Conv" means convolution. "IN" is the abbreviation of instance normalization. "DW" means depth-wise. "C" in the circle represents concatenation. "L2 Norm" is token-wise L2 normalization. The spatial-wise sum is to calculate the summation of all the tokens in Q. Row-wise or Col-wise sum is to calculate the summation of tokens in the same row or column in K. Token selection is to select rows and columns from a matrix with indexes for referring. "×" in the circle represents matrix multiplication. The dotted box on the right represents operations made in a single head of DPSA.

the cost of the contribution measurement can be sharply reduced thanks to the distributive property of the vector inner product.

$$Score_r = \sum_{i=0}^{N} \sum_{j=0}^{W} q_{ik}^T k_{rj} = \left( \sum_{i=0}^{N} q_i \right)^T \left( \sum_{j=0}^{H} k_{rj} \right), \quad r \in \{1...H\}, \tag{3}$$

$$Score_c = \sum_{i=0}^{N}\sum_{j=0}^{H} q_{ik}^T k_{jc} = \left(\sum_{i=0}^{N} q_i\right)^T \left(\sum_{j=0}^{W} k_{jc}\right), \quad c \in \{1...W\}. \tag{4}$$

As in the formula with query and key for each row and column reduced by sum, it can score each of the row and column terms based on the contribution of data. Only tokens with selected row and columns are remained and others are all pruned. As in the formula with query and key for each row and column reduced by sum, it can score each of the row and column terms based on the contribution of data. Only tokens with selected rows and columns are remained and others are all pruned.

$$Index_r = \text{ArgMaxScore}(Score_r)[:N_s], \tag{5}$$

$$Index_c = \text{ArgMaxScore}(Score_c)[:N_s]. \tag{6}$$

The Ns is a hyperparameter that determines how many selections can be done based on the rank. In this experiment, half of height and width are selected. As a result, the selected key and value is changed with Ns at each row and columns with Ks (pruned key) and V (Pruned value). The DPSA computation is as follows.

$$DPSA(X) = \text{concat}_{i=1}^{N_h}[\text{SparseAttention}_i(X_i)]W, \tag{7}$$

$$\text{SparseAttention}(X) = [\text{Softmax}(QK^T)]V_s. \tag{8}$$

By using this pruning process, the computation complexity becomes O(NNC) to O(NNC/4) for this experiment.

### 4.2 temporal embedding

Temporal embedding is done with 2 stages, random projection and a temporal embedding process. Random projection is done for memory limitation. Using frame embedding with 224x28x28 exceeded my memory limitation, and reducing it to 1000 data is also the data I can run. But based on the Johnson-Lindenstrauss Lemma distance equation, the reducing the data from 224x28x28 to 1000 still have 13.5 percent of error bound with original data. Still it would be good to use whole data, but for the memory limitation, 13.5 percent error bound can remain the meaningful original data, and context analysis is already done with ITTR, thus, still possible to understand the overall difference between the frames. The result will be compared in experiment parts. By applying even position to sin and odd position to cos at each frame, the embedding can provide data to the transformer for analyzing the relationship based on the frame difference. The following equation explains the process. PE is positional embedding.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \tag{9}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right). \tag{10}$$

This result now can be computed and provide weight based on the relationship between each frame in the transformer.

### 4.3 Process

As a result, in overall, the CNN encoder can now analyze each image frame and ITTR can provide the complex context in the feature. The temporal transformer can learn importance to each frame, and add to the result of ITTR. This result has both context of single frame and frame to frame relationship. Based on the result of data, the upsampling occurred. This experiment is for broader understanding of how many features and context the model can capture.

## 5 Experiments / Results / Discussion

The foundational experiment design integrates a hybrid loss function, combining L1 loss with Generative Adversarial Network (GAN) loss. The GAN loss framework encompasses a three-stage Convolutional Neural Network (CNN) downsampling process, amounting to a six-stage CNN analysis for the assessment of image recognizability by computational standards. For the critical task of comparing the generated image with the reference image, two regression metrics were evaluated: L1

loss and L2 (Mean Squared Error, MSE) loss. While MSE loss provides a broader overview of error distribution and demonstrates resilience against outliers, L1 loss exhibits higher sensitivity to outliers. The selection of the loss function is inherently aligned with the model's objective; in this case, the aim is to synthesize an image of the forthcoming frame that retains critical features for predictive analysis. Consequently, L1 loss was chosen to quantify the discrepancy between the generated and reference images, supplemented by GAN loss to ensure the synthetic image attains a computationally discernible quality. The learning rates were set at 0.01 for L1 loss and 0.001 for the discriminator (GAN) loss, as determined by empirical results from varied learning rate trials.

The experimental protocol stipulated a batch size of 8, with the dimension of sequential embedding reduced to 1000 due to memory constraints. The dataset comprised authentic urban driving scenarios extracted from YouTube, from which 8000 instances of 5 consecutive frames were randomly selected. This sample size was informed by preliminary experiments to determine an optimal balance for model regularization, ensuring the generation of qualitatively moderate images.

The evaluation encompassed three model configurations: a combined ITTR and sequence transformer model, an ITTR-only model, and a hybrid model incorporating UNet principles alongside ITTR and sequence transformer. The comparative analysis revealed that the integration of ITTR with sequence transformer significantly improved fidelity to the reference image, particularly in capturing dynamic elements, compared to the ITTR-only model, which maintained stable performance for static features but faltered with moving objects.

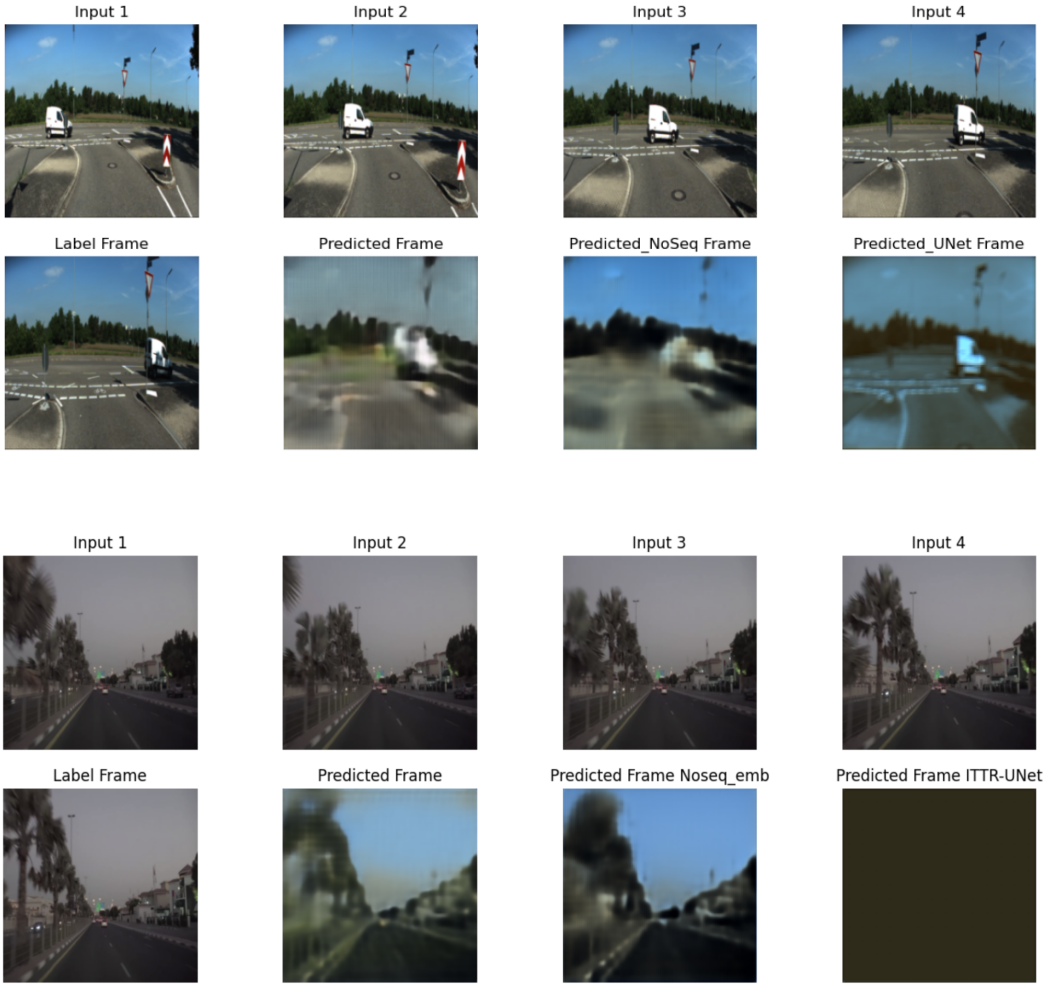


Figure 4: Comparing with dataset with KITTI 360(upper) and YouTube dataset(bottom), the prediction model with sequence embedding could catch more data with changed of input frames.

Table 1: Loss result

| model | train loss | valid loss | test loss |
|---|---|---|---|
| ITTR-Seq | 0.052 | 0.053 | 0.058 |
| ITTR-NoSeq | 0.059 | 0.060 | 0.072 |
| ITTR-UNet | 0.12 | 0.15 | 0.37 |

Furthermore, disparities were observed across different datasets. Notably, models trained on the YouTube dataset demonstrated superior regularization and exhibited lower loss metrics compared to those trained on the KITTI360 dataset. This variation underscores the impact of dataset characteristics on model performance, highlighting the YouTube dataset's efficacy in facilitating more robust model training outcomes.

# 6 Conclusion / Future Work

## 6.1 Conclusion

The integration of the Image Transformer (ITTR) technique offers the potential to enhance computational efficiency through its pruning algorithm, dynamically allocating weights to images to refine the attention mechanism. This selective attention not only conserves data but also enhances the relevance of the processed information. Furthermore, the incorporation of temporal embedding empowers the model to comprehend the sequential relationship between frames, thereby enriching the dataset with a deeper temporal analysis. Experimental outcomes indicate that while the KITTI360 dataset is specifically curated for visual training in automotive contexts, models trained on more recent YouTube video datasets from 2020 to 2022 exhibited superior regularization. This discrepancy is attributed to the contemporary nature of the YouTube data, despite the KITTI360's comprehensive road feature dataset. The latter's older data collection timeframe (over a decade ago) contrasts with the YouTube dataset's more current and varied content, which, although not exclusively focused on road conditions, leads to improved image quality through enhanced generalization.

## 6.2 Future Work

This research was motivated by the objective of developing time-efficient computational models, utilizing personal computing resources and Google Colab due to constraints in computational power and time. The exploration of the full spectrum of hyperparameters and datasets was limited under these conditions. Future enhancements could address two key areas: the refinement of random projection and the optimization of CNN encoding/decoding processes. The sequential transformer, pivotal for its role in contextualizing temporal relationships, could benefit from larger-scale random projection or complete utilization of original data to minimize the 13.5 percent error margin induced by current memory constraints. Furthermore, the architecture of the CNN encoder/decoder, particularly the balance between downsampling and upsampling, warrants further investigation to ascertain the optimal data retention for comprehensive frame analysis. Such explorations were restricted by the existing memory limitations and project timelines.

In term of algorithm, in my opinion, the computation of the model can be more proved with channel term of pruning. In other words, add the addtional pruning process in channel term, and do the additional pruning. The whole process can be faster, and by combining the sequential embedding and transformer inside of HBP, it can analyze more reduced but weighted data in the model.

Moreover, the prospect of employing a larger, higher-quality dataset presents an appealing avenue for future research. The current methodology involves downsizing images from 1280x720 to 224x224 pixels, which inevitably omits finer details. The ability to process higher resolution images could significantly enhance the model's capacity to discern intricate features within the data, thereby improving the overall accuracy and efficacy of the predictive framework.

# References

Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. 2015. Deepdriving: Learning affordance for direct perception in autonomous driving.

Sanjoy Dasgupta. 2013. Experiments with random projection.

Foivos I. Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. 2020. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114.

Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. 2018. Learning to drive in a day.

Wenjie Luo, Bin Yang, and Raquel Urtasun. 2020. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net.

Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. 2018. Multinet: Real-time joint semantic reasoning for autonomous driving.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

ChaoQiang Zhao, QiYu Sun, ChongZhen Zhang, Yang Tang, and Feng Qian. 2020. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, 63(9):1612–1627.

Wanfeng Zheng, Qiang Li, Guoxin Zhang, Pengfei Wan, and Zhongyuan Wang. 2022. Ittr: Unpaired image-to-image translation with transformers.