Bird's-Eye-View from Multiple Vehicle Camera Images

Ayush Goel
aygoel@seas.upenn.edu

Ojas Mandlecha
ojasm@seas.upenn.edu

Renu Reddy Kasala
renu1@seas.upenn.edu

***Abstract*—In the realm of computer vision, there's a pressing need to transform surveillance and on-road camera perspectives into a bird's-eye view, crucial for applications like traffic management and autonomous driving. Our proposed methodology addresses this challenge by leveraging deep learning architectures to capture spatial relationships and depth cues in on-ground images. This approach focuses on understanding temporal consistencies in multi-view images, enhancing top-down transformation accuracy. To overcome limitations of previous techniques, our methodology targets higher accuracy in object placement and spatial consistency.**

Our comprehensive methodology amalgamates images from multiple vehicle-mounted cameras with synchronized timestamps for temporal congruity. Meticulous preprocessing ensures spatial and temporal alignment, vital for eliminating discrepancies. Leveraging the CARLA dataset, we employ Segformer for semantic segmentation, enhancing spatial consistency. A U-Net backbone processes segmented images, and together with a transformer creates a cohesive bird's-eye view. The proposed network architecture ensures contextual awareness and addresses spatial inconsistencies. This novel approach to semantically segmented bird's-eye view showcases the effectiveness of our methodology, validated extensively using CARLA's diverse dataset, ensuring its robustness and applicability in real-world scenarios.

***Index Terms*—BEV, Transformer, Segmentation, Muticamera setup**

I. Introduction

In the ever-evolving realm of computer vision, there exists a growing imperative to effectively translate surveillance and on-road camera perspectives into a bird's-eye view. Conventionally, prevailing techniques have relied upon basic projection transformations, grappling with issues of distortion and a lack of depth comprehension. Historically, approaches have relied heavily on simple projection transformations, which often suffer from distortions and lack of depth understanding. These methods, while foundational, fail to consider the spatial hierarchies and relationships between objects in a scene. By incorporating deep neural networks, our approach aims to capture intricate spatial hierarchies and inter-object relationships. The model's outcomes serve as valuable contributions to a range of critical functionalities in the context of autonomous vehicle systems, including lane tracking, mapping and localization, motion and path planning, as well as providing essential assistance in parking maneuvers.

II. Related Works

In the dynamic field of computer vision, perspective transformation remains pivotal for applications ranging from autonomous driving to urban planning. Several pioneering works have laid the groundwork in this domain, each contributing unique methodologies and insights.

The approach presented in "Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D" [1] hinges on a three-fold process. A distinct advantage of this method lies in its adaptability to diverse camera configurations and its adeptness at managing occlusions. Nonetheless, the technique may grapple with representing intricate scene details.

"Multi-source Domain Adaptation for Semantic Segmentation" [2] diverts its focus to adaptive semantic segmentation across varied domains. By bridging multiple source domains, it ensures heightened generalizability to previously unseen domains. However, it is occasionally challenged by the necessity of extensive domain-specific knowledge.

"BEVFormer" [3] ventures into the transformative power of transformers for perspective shifts. It seamlessly translates standard images into bird's-eye view representations, banking on the self-attention mechanism inherent to transformers. This comes at the price of computational intensity, often associated with transformer architectures.

Lastly, the Inverse Perspective Mapping (IPM) [4] method, employs geometric transformations to morph camera-centric images to a top-down perspective. It however encounters hurdles with scenes that present varied depth or occlusion challenges.

Our proposed methodology synthesizes these foundational techniques. Inspired by the flexibility of "Lift, Splat, Shoot," we've engineered our model to accommodate a spectrum of camera configurations. Drawing adaptive learning cues from "Multi-source Domain Adaptation," we've optimized our method for diverse scene interpretations. Integrating the multisensory fusion essence of BEVFusion [5], our approach has been refined for real-time applications. Finally, augmenting the geometric principles of IPM with deep learning, we've incorporated a nuanced understanding of depth variations and occlusions.

By understanding the strengths and addressing the limitations of previous works, our project endeavors to push the boundaries in camera perspective transformations.

III. Dataset

In the realm of autonomous driving, several open-source datasets have been meticulously curated to cater to a myriad of

applications, including the generation or adaptation for bird's-eye view (BEV) perspectives.

In our research, we leveraged the CARLA simulator dataset as a foundational component of our project. This dataset encompasses comprehensive perspectives, including left, right, front, and rear views captured by a vehicle-mounted camera, as well as an aerial top-down view from a drone's perspective. Initially, we considered utilizing either the Waymo Open Dataset or the KITTI dataset for our research. However, these datasets did not provide the specific combination of four camera views mounted on a vehicle, along with a top-down view from above, which was a crucial requirement for our project. As a result, we opted for the CARLA simulator dataset to fulfill our dataset needs.

The chosen dataset utilizes five cameras with known intrinsics,

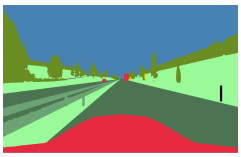

Fig. 1. Example front view image

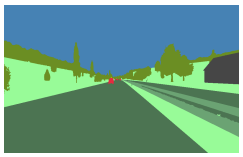

Fig. 2. Example rear view image

Fig. 3. Example left view image

Fig. 4. Example right view image

Fig. 5. Example top-down view image

making it an ideal fit for our specific application. The dataset includes 46,000 images for each viewpoint, and the ground truth is derived from a top-down camera view. In addition, comprehensive details of camera intrinsics and extrinsics, encompassing parameters such as fx, fy, px, py, Euler angles (yaw, pitch, roll), along with camera translations relative to the world frame, are explicitly known for each camera within the setup.

IV. Methodology

In our comprehensive methodology, we propose the amalgamation of images obtained from multiple vehicle-mounted cameras, capturing scenes around the vehicle. We place a strong emphasis on synchronized timestamps to ensure temporal congruity. A meticulous preprocessing phase follows, designed to affirm both spatial and temporal alignment of the gathered data. This phase not only ensures consistency but also eliminates discrepancies within the dataset. Recognizing the inherent design of convolutional layers, which operate locally, and the imperative to adaptively handle images from varied viewpoints, we introduce a multi-modal neural network. Tailored precisely for this purpose, the network distills relevant features from the multitude of images emanating from different camera angles. This adaptive approach is crucial for handling the challenges posed by varying perspectives in the dataset.

Our methodology leverages the CARLA dataset, comprising images from left, right, front, and rear-view cameras. We have also developed the pipeline for real camera images employing a Segformer for segmentation on each image.

To mitigate the spatial disparities introduced by localized convolution operations, our methodology strategically incorporates Inverse Perspective Mapping (IPM). This crucial technique is employed to improve spatial consistency between input and output images. Recognizing the presence of multiple camera setups, we make a deliberate decision to create individual heads exclusively dedicated to executing IPM. Each head is precisely aligned to cater to a specific camera view, ensuring accuracy in the transformation process.

After the segmentation phase, we leverage the IPM to derive the homography, which serves as our ground truth. This step allows us to transform images from individual camera viewpoints into a unified top-down perspective, thereby enhancing spatial consistency in the overall scene representation. This meticulous approach not only addresses the challenges introduced by localized convolution operations but also contributes to a more cohesive and accurate representation of the scene.

To generate the Birds Eye View (BEV), a U-Net backbone is employed within our proposed network architecture. It plays a pivotal role in processing segmented images from multiple viewpoints ensuring a cohesive integration of information, and addresses spatial inconsistencies thereby providing a comprehensive BEV representation. The detailed information about architecture is as follows.

A. U-Net Backbone [7]

The centerpiece of our methodology is the integration of the renowned deep learning architecture, **U-Net**. This architecture is lauded for its prowess in tasks such as semantic segmentation. By leveraging U-Net's structured convolutional encoder and decoder mechanisms, we extract nuanced features from the input data, pivotal for generating an accurate BEV. Furthermore, given our multi-camera setup, the U-Net ensures that images from the four vantage points are cohesively processed, and spatial inconsistencies are adeptly addressed.

B. Network

Our proposed network architecture represents a novel approach to semantic segmentation, featuring a multi-input encoder-decoder framework enriched with a spatial transformer module. This design is crafted to enhance the model's ability to comprehend diverse input images effectively. The encoder component of the network is tasked with extracting hierarchical features from the input images. This hierarchical

representation is crucial for capturing intricate patterns and details within the visual data.

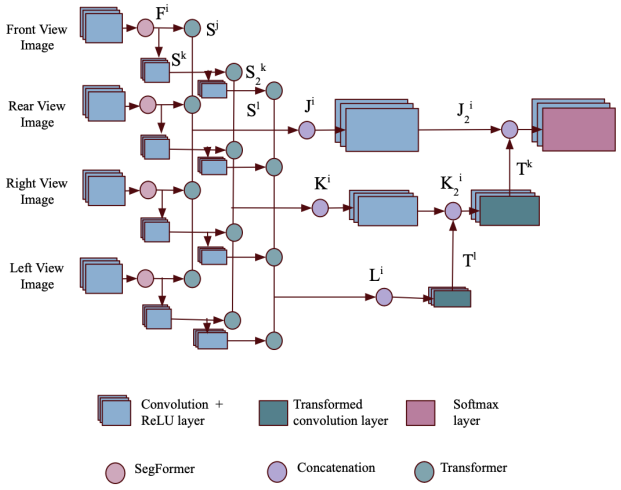

Fig. 6. Network Architecture

Incorporating a transformer module further elevates the network's capabilities by dynamically adjusting spatial relationships within the feature maps. This adaptability proves invaluable across a spectrum of input scenarios, allowing the network to account for varying spatial orientations present in complex visual scenes. The transformer's role in refining the spatial relationships contributes significantly to the network's contextual awareness and overall performance. The concate module is a pivotal component that fosters enhanced context awareness by facilitating the joint processing of both spatially transformed and original feature maps. This integration ensures that the network leverages the synergies between the spatially transformed representations and the inherent contextual information within the original feature maps. This collaborative processing strengthens the network's ability to make informed decisions during semantic segmentation.

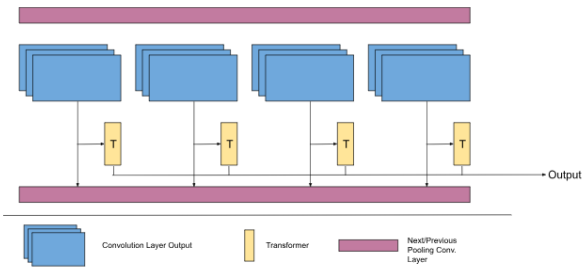

Fig. 7. Transformer Block.
The output of transformer blocks are integrated into a single feature map at every level

The decoder, mirroring the encoder architecture in reverse, plays a crucial role in reconstructing the final segmented output. The gradual upsampling and integration of features from the fused encoder layers contribute to the refinement of the segmentation results. This bidirectional architecture, coupled with the transformer's capabilities, demonstrates a robust approach to top-down view construction, particularly in scenarios with varying spatial orientations.

Additionally, we used Segformer for obtaining semantically segmented images of CARLA dataset, further enriching our network's ability to comprehend and interpret complex visual scenes.

The CARLA input (Left, right, rear or front) image is passed through the SegFormer model and the output is denoted as $F_p^{(l)}$ where l is the layer and p denotes the position (ex. right, left, rear, front).

$$F_p^{(l)} = SegFormer(Real\ Image) \tag{1}$$

$F_p^{(l)}$ is then downscaled by passing through a convolutional layer followed by a Rectified Linear Unit (ReLU) activation function.

$$s_p^{(l)} = Downscale(F_p^{(l)})$$

The output of this layer is then pass to transformer layer and the next convolution layer.

$$t_p^{(l)} = Transformer(F_p^{(l)})$$

$$s_p^{(l+1)} = Downscale(s_p^{(l)})$$

$s_p^{(l+1)}$ is passed to the transfomer and convolution of the next layer. This constitute as one layer for one camera. There are 3 such layers for each camera. The output from each of the transformer (for each camera) for every layer is concatenated and then convoluted (Ref fig 7). In last layer of encoder ends with a transformer for each camera.

$$J^{(l)} = Conv(Concat(t_{right}^{(l)} + t_{left}^{(l)} + t_{front}^{(l)} + t_{rear}^{(l)}))$$

These outputs are then upscaled by passing through a transformed convolutional layer and then added to the subsequent layer above it.

$$d^{(l)} = Upscale(J^{(l)})$$

$$J^{(l)-1} = d^{(l)} + J^{(l)-1}$$

This forms the decoder of our model which gives us the required output at the end. This can be seen easily visualized in the Fig. 6.

This seamless integration combines the strengths of our proposed architecture with cutting-edge segmentation techniques, making it well-suited for applications demanding a comprehensive understanding of diverse and intricate visual data.

D. Evaluation and Testing

An intrinsic part of our methodology is the rigorous testing phase. This ensures the model's robustness and reliability

Fig. 8. Example real front view image

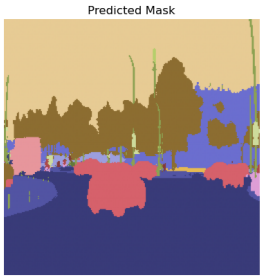

Fig. 9. Example real rear view image

Fig. 10. Example segmented image of front view

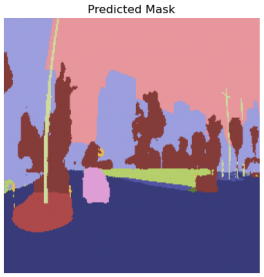

Fig. 11. Example segmented image of rear view

when introduced to real-world, unseen data.

V. Setup

A. Dataset Post-Processing:

In our autonomous driving research, we utilized the CARLA simulator dataset, containing 46,000 images for each viewpoint. Due to computational constraints, we opted for a subset of 50,000 images, evenly distributed with 10,000 images per view. Ground truth, obtained from a virtual drone camera, includes a Bird's Eye View (BEV) image centered above the ego vehicle.

For model training, we allocated approximately 70% of the data, ensuring the model learns patterns and features effectively. A 15% subset was reserved for testing to evaluate the model's performance on unseen data, while the remaining 15% was designated for validation. This validation set played a crucial role in fine-tuning hyperparameters and preventing overfitting during the training process.

B. Training:

The model was trained on the images of size 650px x 450px. A pipeline for the images to pass through pretrained state-of-the-art Segformer-B5 model to get semantically segmented images is also designed to be used on real data. Images are then converted to one-hot encoding. These one hot encodings can be changed as per need to get as detailed output as possible. The loss function is adapted to assign weights to semantic classes based on the logarithm of their relative frequencies. This adaptation helps with imbalance in the data collected. The optimizer used was Adam. Since we are dealing with classes, the logical loss fucntion is cross-entropy loss.

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p_i)$$

In our project, the model was trained on an AWS g5.4xLarge instance. This instance type, is equipped with NVIDIA A100 Tensor Core GPUs, 24GiB GPU Memory, 16 vCPUs, 600GiB CPU Memory. This robust instance type provided the necessary resources to efficiently train and optimize our neural network, enabling accelerated progress. The best results achieved were using the following hyperparameters:

TABLE I
HYPERPARAMTERS

Name	Value
Batch size	8
Learning Rate	5e-5
Betas	(0.9, 0.99)
Epochs	30
Weight Decay	1e-6

C. Evalution:

The model is validated on the dataset, ensuring a balance in semantic class distribution. Testing was executed on unseen data to simulate real-world application. The cornerstone of our evaluation is the Intersection-over-Union (IoU) score. IoU offers an insight into the model's precision by quantifying the overlap between the predicted and the actual ground truth regions for each class. At the end, we have calculated Mean Intersection-over-Union (MIoU) which aggregates these scores, giving an overall metric of how adeptly the model predicts across multiple semantic classes.

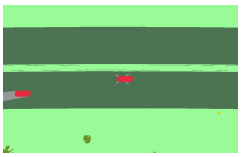

Fig. 12. Example predicted BEV image

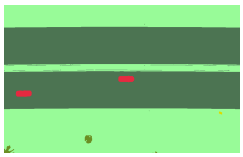

Fig. 13. Example ground truth image

VI. Results and Discussion

Our method variations are meticulously assessed, with a focus on the model, which uses uNet as baseline, as the primary approach. The standard homography image obtained through IPM serves as the baseline for comparison. MIoU scores on the validation set indicate the superior performance of our model, achieving 55.31%. Ablation studies, removing IPM from our approach, reveal the significant performance boost provided by our methodology compared to the homography baseline. Class IoU scores further validate the effectiveness, particularly in predicting large-area semantic classes. Despite challenges in predicting smaller objects like bikes and persons due to their infrequent occurrences in the

dataset, our dataset showcases improvement, leveraging raw camera images. To improve the result in spite of imbalance, we included logarithmic weights to each class in our loss functions.

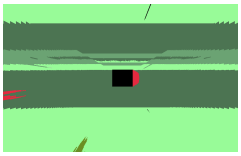

Fig. 14. Example baseline(1) IPM BEV image

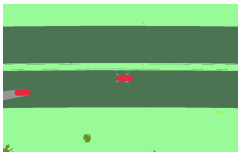

Fig. 15. Example our final BEV image

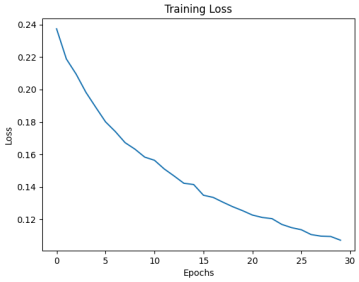

Fig. 16. Training Loss

Fig. 17. Validation Loss

Semantically segmented images play a pivotal role in dimensionality reduction, focusing computational efforts solely on relevant spatial information essential for generating accurate Bird's Eye View (BEV) representations. By discerning and classifying objects at a pixel level, the model selectively processes significant visual features, effectively minimizing irrelevant details and enhancing the efficiency of the subsequent BEV generation process. The BEVs generated by our model demonstrate the potential robust predictions in real-world scenes as well. Despite the fixed IPM transformation assumption, our models exhibit promising results, indicating potential refinement with dynamic transformation adjustments based on vehicle dynamics in future implementations. Below is MIoU scores on the validation dataset in comparison with our model

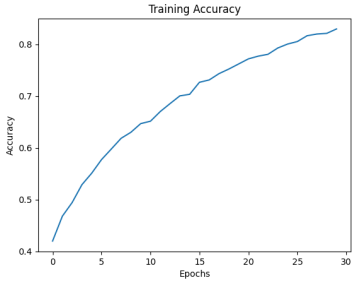

Fig. 18. Training Accuracy

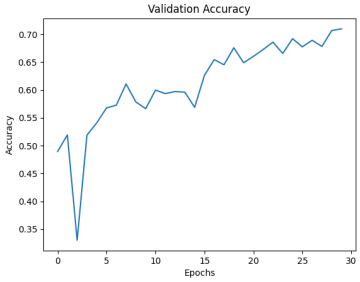

Fig. 19. Validation Accuracy

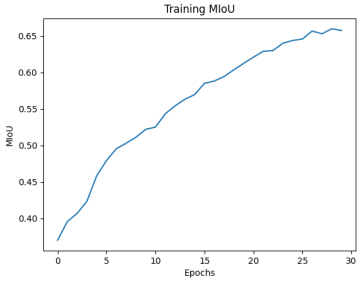

Fig. 20. Training MIoU

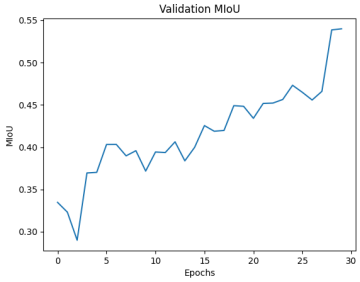

Fig. 21. Validation MIoU

TABLE II
MIoU Scores

Model	MIoU (%)
Our Model	55.31
Homogrpahy	31.19

VII. Conclusion

In conclusion, our proposed methodology adeptly transforms images from multiple vehicle-mounted cameras, miti-

TABLE III
CLASS IOU SCORES

Name	IoU (%)	Log(freq)
Road	92.28	0.913
SideWalk	88.18	2.24
Car	66.39	4.78
Person	9.43	10.47
Flora	95.82	1.02
Truck/Bus	55.36	7.56
Other	45.10	3.64
Mean	55.31	

gating errors stemming from the inaccurate flatness assumption inherent in Inverse Perspective Mapping.

We successfully implemented a regular model on simulated data, gaining valuable insights into the challenges posed by multiple vehicle-mounted cameras. Subsequently, we developed a real-life data pipeline using Segformer for semantic segmentation. Subsequently, we developed a real-life data pipeline using Segformer for semantic segmentation which can be potentially used on real-world data.

Our proposed methodology adeptly transforms images captured by multiple vehicle-mounted cameras into bird's eye view (BEV), effectively mitigating errors associated with the inaccurate flatness assumption inherent in Inverse Perspective Mapping. Leveraging datasets and employing an input abstraction to semantically segmented representations, our approach seamlessly extends to real-world data without necessitating manual labeling of BEV images.

Looking towards the future, our work holds significant potential for further advancements. We aim to explore enhancements that address finer details in semantic segmentation, refining the network architecture for more nuanced feature extraction. Additionally, extending our methodology to diverse real-world scenarios is a crucial next step. We also plan to investigate the integration of dynamic transformations to accommodate varying camera poses in real-world environments, enhancing the robustness of our approach.

Furthermore, the incorporation of depth information is a promising avenue for improving the perception capabilities of our methodology. By integrating depth data, our approach can achieve more accurate scene understanding, aiding in the precise localization and identification of objects. This additional dimension of information has the potential to significantly enhance the robustness and reliability of our method across various complex real-world scenarios. As we continue our research, these future directions will contribute to the evolution and refinement of our proposed methodology.

REFERENCES

[1] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D" in ECCV, 2020.

[2] S. Zhao et al., "Multi-source domain adaptation for semantic segmentation," in Proc. NeurIPS, 2019.

[3] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," arXiv preprint arXiv:2203.17270, 2022.

[4] H. Mallot, H. Bulthoff, J. Little, and S. Bohrer, "Inverse perspective mapping simplifies optical flow computation and obstacle detection," Biological Cybernetics, vol. 64, no. 3, pp. 177–185, 1991.

[5] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "BEVFusion: A simple and robust lidar-camera fusion framework," arXiv preprint arXiv:2205.13790, 2022.

[6] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, C. Zhifeng, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in CVPR, 2020.

[7] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in CVPR, 2019

[8] A. Laddha, S. Gautam, S. Palombo, S. Pandey, and C. Vallespi-Gonzalez, "MVFuseNet: Improving end-toend object detection and motion forecasting through multi-view fusion of lidar data," in CVPR Workshops, 2021.

[9] Z. Qin, J. Chen, C. Chen, X. Chen, and X. Li, "UniFormer: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view," arXiv preprint arXiv:2207.08536, 2022.

[10] T.-Y. Lin, P. Dollar, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," CVPR, 2017.

[11] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in ICLR, 2020.