Transformer-Based 3D Reconstruction from Single-view Images

Harshal Shrimali
MS in Data Analytics
San Jose State University
San Jose, California

Raveena Kagne
MS in Data Analytics
San Jose State University
San Jose, California

Samruddhi Gawande
MS in Data Analytics
San Jose State University
San Jose, California

Vishal Yeole
MS in Data Analytics
San Jose State University
San Jose, California

***Abstract*—One of the great challenges in computer vision is the reconstruction of 3D scenes from a single 2D image due to the absence of depth information and occluded parts of the scene. Most approaches are based on generating 3D models from multiple views or using specialized devices such as LiDAR, making them impractical in scenarios with limited visual data. In this project, a new pipeline is presented that incorporates state-of-the-art models with transformers and generative AI to resolve the limitations. Given only one RGB image, the framework deploys vision-language models including BLIP and LLaVA that help generate semantic descriptions to improve contextual understanding. The segmentation level implemented with OneFormer makes it possible to detect the actual elements of the image such as roads, buildings, and other objects. The pipeline scope out the field of view and in a subsequent step employ generative models, namely Stable Diffusion SD 3.5, to fill in areas that have not been covered, thus completing the scene. The depth maps generated, by MiDAS or Depth-Anything models, are the input depth estimation algorithms for creating point clouds in Open3D. These point clouds are turned into meshes to obtain advanced 3D scaffolds. The proposed approach was tested using the Cityscapes dataset, and the results demonstrated an improved level of structural fidelity with an SSIM score of 0.48 leading to improved visual coherence. This method can be use in self-driving cars, urban design, AR or VR environments.**

***Index Terms*—3D Reconstruction, Single-View Images, Transformers, Depth Estimation, Gaussian Splatting, Urban Planning**

I. Introduction

The computer vision of reconstructing 3D structures from single views is one of the hardest problems which can be useful in places such as self-driving cars, city development, or in augmented reality. Most popular and traditional methods of 3D reconstruction, which include stereo vision, multi-view geometry, or specific sensors such as LiDAR, are convenient and effective, but often include stereo images or images taken with calibrated devices, and require much computational power to perform these tasks. Still, they have major limitations when it comes to scaling, especially in real-world situations when only one image is available (Han et al., 2019; Henderson Ferrari, 2019). Recent years have seen great progress towards one-view 3D reconstruction due to CNNs that are combined with large datasets that allow them to apply spatial and geometric priors, significantly speeding up the processes. However, because of design, CNN-based methods become difficult to transfer to complex scenes, efficient computation, and scalability limiting them in most practical applications. (Fu et al., 2020; Shi et al., 2021). Transformer architectures, first created to solve problems in NLP, have recently shown promise for 3D reconstruction tasks. This gives them a high ability to reconstruct 3D objects from 2D images (Zou et al, 2024; Wu et al, 2024). Reconstruction of 3D images from single-view images is also known as single-view 3D reconstruction. It poses quite several challenges, for it is ill-posed problems intertwined with depth-fusion, occlusion, and heterogeneous geometry. In recent studies, several solutions have been proposed for these problems using generative modeling decoupled with hybrid 3D representation techniques such as 'Gaussian splatting' and triplane methods to optimize efficiency with precision making the model generation process efficient. VistaDream is another framework that advanced this field by combining diffusion models to obtain multiview consistency to improve the reconstruction of the scenes, making them more coherent (Wang et al. 2024; Zou et al. 2024). For a more complete geometric and visual response, this project proposes a transformer-based framework for reconstructing 3D images via single-view images and embedding more cutting-edge models and technologies. The framework exploits vision-language models such as BLIP and LLaVA for semantic comprehension, OneFormer for spatial partitioning using segmentation transformers and MiDas, and Depth Anything v2 for spatial geometry capturing. To further enhance fidelity, the principles of Gaussian splatting are employed to represent 3D efficiently and mesh generation for structural coherency. These innovations seek to improve current methods and close the disconnect between 2D and 3D representations (Han et al., 2019; Henderson Ferrari, 2019; Wu et al., 2024). This study proposes extensions on urban simulation, self-driving vehicles, and augmented reality. Future research focuses on exploiting multiview consistency approaches and the use of more sensory data such as LiDAR to achieve greater accuracy and robustness (Fu et al., 2020; Zou et al., 2024).

II. Related Work

A. Reconstruction Approach

SyncDreamer resolves the problem of creating multiview-consistent images from a single view input by presenting

a synchronized multiview diffusion framework. Rather than considering new views as separate, SyncDreamer learns the joint probability distribution of all views. The framework incorporates a 3D spatial feature volume reconstruction based on attention mechanisms across noisy views. This approach promotes the use of epipolar constraints and improves consistency by unifying corresponding features across multiple views. SyncDreamer is based on the Zero123 architecture which has been known to offer good generalization and allows different input styles such as sketches or artistic renditions. The synchronized noise predictors used by SyncDreamer always perform the reverse diffusion for all target views at the same time which maintains geometry and appearance consistency. In contrast to distillation-based methods, SyncDreamer does not have complex and time-consuming text inversion and optimization phases. There are many reasons why SyncDreamer can be favored when it comes to 3D reconstruction. It is capable of producing consistent images in multiview configurations, maintaining geometry as well as appearance. Its generalization does not depend on the type of input, for instance, even a hand-drawn picture can be used as input, making it usable for different purposes. Moreover, the concurrent generation technique averts wasting effort. However, due to the nature of diffusion and the difficulty of synchronizing the views, SyncDreamer is very resource-demanding. This also means that it may have some potential. drawbacks for applications where arbitrary scene flexibility are required, as fixed viewpoints can encourage classification. [1] Transforming 2D images into intricate 3D scenes is an ill-posed task which has the downfall of being inconsistent across unseen views. VistaDream offers a two-stage framework to overcome these challenges. In the first stage, an attempt to build a global coarse 3D scaffold is made, which is further extended by zooming out the Field of View (FoV), and later regions are inpainted with models like Fooocus. A depth map is then constructed upon the zoomed-out image, whose purpose is to simply provide a rough 3D geometry. This procedure provides good constraints through disparate views of the image and reduces denting errors in the next warp and inpaint steps. A new MCS algorithm is used in the second stage to enhance the generated novel views. In contrast to previously used methods like Score Distillation Sampling (SDS), MCS reduces the number of views sequenced for generation to one in the reverse diffusion phase, making the resulting 3D scene high-quality, and more integrated than before. VistaDream has several upsides. It avoids additional computational burdens since it does not require the training or fine-tuning of further diffusion models. Also, due to the incorporation of MCS, the view-inconsistency denoising artifacts are eliminated throughout the diffusion process. Besides, it generates scenes with greater quality and structural integrity than the GenWarp and the RealmDreamer-based methods. There are however some constraints that come along with VistaDream. The 3D scaffold's initial points of view have a strong dependency on the depth maps. Inaccurate depth calculations can therefore distort images all over the pipeline. In addition, the refinement stage also sometimes results in a slight degradation in detail during the strengthening of multiview consistency. VistaDream's methodology of scaffolding and multiview consistency algorithms can solve a number of problems during this project, including zooming out, inpainting, depth estimation, and multi-view generation. Because of its global scaffolding and consistency refinement, it can solve geometric inconsistencies and enhance overall reconstruction quality [2].

B. Image Captioning

LLaVA addresses the problem of integrating vision and language models for applications such as image-text generation and multi-modal comprehension. This creative method enhances the usefulness of large pre-trained vision models, such as CLIP, and large language models, such as GPT-3 or T5, by developing a joint multimodal training framework. Due to LLaVA's vision-language alignment strategy, caption generation can be done more precisely since the model takes into account the image content as well as the text description. The model generates using VLMs which are already trained not only to say what objects are in the image but also to tell what the image depicts in detail. Such a method is superior to image captioning models that are based on a task-oriented approach and trained on a limited number of images. LLaVA has a good number of advantages. One of them is its capacity to produce captions that correspond with the features of the images and are also comprehensible in natural language. Furthermore, since LLaVA implements pre-trained models, the time and resources that are normally spent on building an image captioning system from ground up are lessened. It is also a great advantage that the model is able to transfer learn from different tasks which include image captioning and visual question answering (VQA). LLaVA is capable of performing several multimodal tasks because it uses vision and language models without the need to heavily fine-tune each for the individual task. Nonetheless, there are also shortcomings that are associated with the use of LLaVA. One of them is its reliance on large scale pre-trained models thus increasing the computational costs especially in the case of high resolution images and large image datasets. The model can also fail due to noise if the image data used is of a very low quality. There are also instances where the alignment of vision and language models can be very complicating and even lead to performance issues when attempting to carry out specific task requirements that are often too complex in nature for the images used. The advanced visual understanding and captioning capability that LLaVA possesses makes it suitable for the captioning phase of this project. The fact that the Cityscapes dataset features intricate urban images and can be succinctly described with accuracy will be important for understanding future operations like depth estimation and 3D scaffolding context. [3] Another framework that is robust for vision language understanding is BLIP. BLIP's method utilizes a vision encoder which extracts high-level semantics from images that are subsequently used by a language model in generating descriptions. A contrastive

learning model that has been developed for text and image alignment is employed. This ensures that captions generated, describe the contents of the images. This makes it possible for BLIP to generate captions which not only describe the objects in the images but also more advanced relationships and contexts where the objects are placed. The main benefit of BLIP is that it is generally applicable. It is made to work well on a variety of tasks, such as visual question answering (VQA), image captioning, and image-text matching, without requiring a lot of task-specific fine-tuning. BLIP can adjust to various multimodal jobs thanks to its flexibility, which also lessens the requirement for specialist training. Furthermore, BLIP successfully aligns text and image representations with the aid of contrastive learning, resulting in captions that are both contextually appropriate and of high quality. BLIP has weaknesses too, even if it possesses many useful qualities. One defining problem is the use of contrastive learning which, even if it works well, may not encompass all the complexities of interaction between the visual aspects and textual ones, leaving behind finer or more domain specific details. Furthermore, although BLIP yields better efficiency than a few other models, its pretraining stage still is very expensive in terms of computing infrastructure and the amount of data with labels, hard for sparse datasets or low-resourced settings [4].

C. Segmentation

Traditionally, the image segmentation problem including semantic segmentation, instance segmentation, and panoptic segmentation has been approached as three different problems with three models for each of the tasks. These are usually separate models and so they are often trained independently, which increases computational costs and does not support scaling or variety. To tackle these weaknesses, OneFormer proposes a new multi-tasking framework that can carry out all three segmentation tasks using only one model that has been trained once but for several objectives. This method appears to be more efficient than the prior one as it reduces the amount of complexity and resources required for the training of separate models for every task that is being performed. OneFormer demonstrates state of the art results on all three segmentation tasks owing to its multi-task architecture which makes the model agnostic to each task of segmentation but learns panoptic annotations pertaining to the intersection of these tasks. Combined with generalized task tokens that direct the model during the training phase and a subsequent multi-faceted detection phase, this task conditioned approach allows the model to efficiently perform many segmentation tasks without being over fine-tuned. This mechanism of task-token is significant in making the model task-dynamic, enabling it to meet the specific demands of semantic, instance and panoptic segmentation. Additionally, a query-text contrastive loss is proposed for the purpose of enhancing the model's ability to distinguish between tasks. Such loss may assist in linking object queries to textual representations and as a result increase the capacity of the model for more diverse categories while making sure the predictions remain task specific. The architecture of OneFormer consists of a transformer-based pixel decoder and a multi-scale deformable transformer that increases the accuracy of prediction and extraction of features by the model. It also allows the use of advanced backbones like ConvNeXt and DiNAT, which help achieve better efficacy and scalability across datasets like ADE20k, Cityscapes, and COCO. Even though Mask2Former has been trained separately for each task with much more resources, OneFormer outperforms such specialized models on benchmark datasets. This multi-task approach not only enhances efficiency but also guarantees satisfactory results for a wide variety of real-life use cases. Still, OneFormer has its disadvantages. The intricacy of multi-task learning can make it difficult to tune the model for specific tasks. Also, OneFormer is transformer-based architecture, which is quite computationally intensive thus its use in low resource settings may not be very practical [5]. All the qualities of OneFormer make it an ideal fit for complex tasks that require both strong performance and computational resource management such as urban scene segmentation, autonomous driving, and augmented reality.

D. Zoom Out and Inpaint

Stable Diffusion is the newest method to combine generative tasks in which images are generated using diffusion processes. It is based on the principles implemented in the family of denoising diffusion probabilistic models (DDPM) models and converts random noise into images in a sequence of induced denoising steps based on learned noise maps. Latency space representation is the main improvement of stable diffusion which allows data to be manipulated in compressed states. A major increase is in the quality of the output while at the same time a significant amount of the resources needed are saved when doing the processing. The model in addition uses the adaptive noise schedules and attention mechanisms to enhance the rate of training convergence and concentrate on critical parts of data. Some of the major tasks performed are text to image generation, inpainting, and even image super-resolution all within short periods of time while allowing users to provide conditional inputs such like existing images or textual descriptions. The availability of its open-source code has also ensured that it can be widely deployed in industries and applications that are creative in nature. Nonetheless, Stable Diffusion is not devoid of shortcomings. It employs models such as CLIP that is known for its text conditioning, which adds to its computational and storage requirements. Also, although the latent-space approach allows for a more efficient process, it might cause artifacts to appear or high details to be lost in certain instances. The model's efficiency also has a very strong bond with the amount and diversification of the training data used for this purpose [6]. Generative models have played a significant role in the improvement of creative processes, facilitating high-quality image and video creation. With regards to this, the paper by Runway ML explores a new generative framework developed in the context of latent diffusion models (LDMs). LDMs solve the drawbacks associated to the conventional generative models such as their focus on computation

and the reliability of scale-dependent findings by employing a latent approach. This approach considerably alleviates the computational burden of these outputs while preserving their accuracy and resolution. The tools developed by Runway ML, which seek to simplify the access to machine learning tools in creative fields, greatly benefit from such progress. The new framework incorporates diffusion probabilistic approaches to be applied in the transforming of stochastic data back and forth until the optimal output is achieved. To sufficiency latently facilitates the use of the model in occupation of excessive resource. The addition of attention helps improve the quality of the generated images by allowing feature extraction and editing to be much more accurate. Such changes are tuned greatly to Runway ML's tools as they are constantly developed for creating, editing, and transforming images. Furthermore, the open-source nature of the framework makes it easy to access and change, which promotes its use and experimentation. At the same time the scalability of the model demonstrates high image synthesis capabilities but with a much smaller number of parameters than other generative models. What's more, this efficiency is in keeping with Runway ML's goals to extend advanced AI functionalities even to non-developers through user-friendly interfaces and less demanding processes. The novel system, however, still has some drawbacks. The other important limitation is the dependence on a large volume of high-quality training datasets for application in data-deficient areas. In addition, although the latent-space method improves the performance, it can introduce small noise in rather sophisticated or dynamic visual scenes. [7]

E. Depth Map

Depth estimation through monocular images is a challenging task with various applications including autonomous navigation, 3D reconstruction as well as augmented reality. Other than the quality of the model, the diversity of the training datasets is as much if not more, important. While traditional depth estimation methods utilize a vast amount of data, they have a tendency to overfit to the characteristics of the particular data collection environment or types of images, thus being unable to perform well to entirely new data. MiDas, as first introduced by Ranftl et al., deals with these problems by presenting a new method that combines several datasets at the training stage. This approach makes it possible to learn robust models even when there are discrepancies in the annotations, scales, or depth ranges across the datasets. MiDias has a robust training objective that is resistant to these factors and utilizes a well-constructed multi-objective learning framework for effective integration of data from multiple sources. One of the greatest benefits of MiDaS and what makes it unique is its cross-dataset transfer effectiveness, which permits evaluating the model on unseen datasets. The results prove that MiDaS strengthens the existing models and establishes a new standard in monocular depth estimation. It is worth noting that this technique is most suitable for practical cases where one has to estimate the depth in many different scenes accurately. However, it has some limitations. Due to the nature of the approach, availability and quality aid of multiple training datasets are principal requirements, which might not be always available or might need extensive preprocessing to fit into the training requirements of MiDas. Furthermore, though MiDas makes progress in the generalization area, the significant amount of graphical processing power associated with many datasets and sturdy loss function training might limit training speed and scalability [8].

Depth Pro brings important progress in monocular depth estimation, as it attempts to solve several key issues preventing existing models from practical deployment scenarios. This model created by Bochkovskii et al. is able to produce highly detailed and sharp depth maps without the need of metadata such as camera intrinsics, a limitation common in previous models [7].

Depth Pro is able to create 2.25-megapixel depth maps multimodally within 0.3 seconds on a standard GPU and generate orders of 75,798 depth maps with the resolution of 672 x 448 within just an hour. This is important because applications like image editing and advanced view synthesis often require Finer details. It has also been shown that the model has performed well in a zero-shot setting across different scenarios producing metric depth maps that are consistent with the absolute scale of the scene without the need for any domain specific tuning. Boundary sharpness in depth map images is also an area that is well catered for through the multi-scale ViTs used in the model, which promotes local detail without losing the global context. Such an approach also benefits from mixed training protocols with real and synthetic data, enhancing the model´s generalization further as well. On the other hand, under the mixed training protocol, some generalization is possible, but the number of completely different scenes can be reached only if there are a large number of different training datasets. This might restrict the performance of the model in less represented areas or environments and this concern is valid scope of concern. However, the need for lots of hardware resources, thanks to the multi-scale ViT and the processing of high-resolution images, seems to limit its deployment in low-resource settings. The complex training process, which is involving real as well as synthetic datasets, is helpful on both accuracy and generalization but may create a challenge on the training since careful adjustments and validation are needed so that overfitting or underfitting is avoided [9].

Depth Anything V2 is an advancement in monocular depth estimation (MDE) that overcomes the challenges and shortcomings encountered by the likes of MiDaS and Depth Pro. This model, developed by Yang et al., effectively synthesizes high-quality depth images from a single input image, allowing it to make impressive advances in zero-shot techniques in complex and diverse contexts. Using both synthetic and vast amounts of real image pseudo-labeled data, Depth Anything V2 aims to add detail and improve the quality of the depth mapping. Unlike previous versions, however, it manages to close the gap between the quality of predictions made on synthetic images and the quality of predictions made in real life through effective modeling techniques. Depth Anything

V2 is a diverse model due to its range of small 25M parameter model to a larger 1.3B parameter version. The parameter range allows for the model to be useful in several practices such as mobile devices as well as high computational tasks. Depth Anything V2 is now capable of outperforming previous models due to its advanced model, which was holistically validated against a novel versatile benchmark aimed at real-world application. This benchmark has a variety of conditions therefore the model will be applicable in most known uses. Although synthetic and pseudo-labeled data allow bypassing limitations of real labeled datasets, they also create an added dependency on the existence of this data and its quality, which might not always be able to cover all real-life scenarios with a proper thoroughness. Perhaps a more complex scenario would be the training process of the model, which involves the use of high-quality synthetic images as well as pseudo labeled images of real objects. This may entail that such models are not easily implementable and perhaps adaptable in environments that are not richly endowed with resources [10].

F. Point Cloud

Originally developed by Intel Lab's researchers Qian-Zhou, Jaesik Park, and Vladlen Koltun, Open3D is an open source software library with real potential for rapid development and processing of 3D data. The library is equipped with both C++ and Python interfaces and allows to work with point clouds, meshes, and RGB-D images effortlessly. Open3D reduces development time and improves efficiency by providing a well-chosen set of data structures and algorithms that simplify the programming needed for complicated 3D data tasks. Open3D is efficient, with a backend built for OpenMP pptimized parallel execution and is suitable for High performance computing environments. It is cross-platform and can be easily built from source due to the low number of dependencies and their strategic placement. There is plenty of documentation and the library is updated often. This means that both academic researchers and those working in businesses are able to implement complicated research with helpful sources. Dealing with 3D data in Open3D seems to be easy, but new users of the software can be confused by its structures and algorithms. There are few dependencies, but the authoring and deployment on various platforms can be problematic for some users with low programming experience [11].

III. Methodology

The methodology is carried out in a modular and systematic way aimed at reconstructing 3D scaffolds from single-view RGB images. Each stage incorporates the cutting-edge models and technologies of the time, hence the integration and the entire flow are smooth and effective as depicted in Fig1 and Fig2.

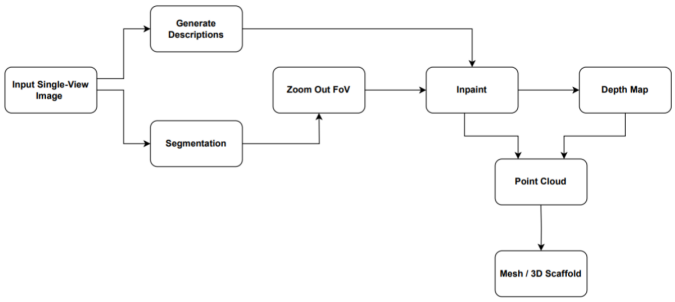

Fig. 1. Proposed framework

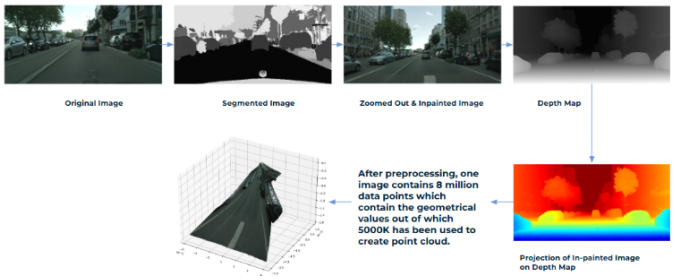

Fig. 2. Representation of Process Flow

A. Data collection

The Cityscapes dataset was created to facilitate dense annotations of urban street-level imagery to promote semantic understanding applications within the computer vision region, more specifically, for self-driving vehicles. The data was harvested in outdoor urban settings in 50 European cities. The cities were selected to incorporate all sorts of factors such as geography, season, weather, architecture, and time of the day, targeting maximum variability. Such variability makes sure that the dataset is incredibly rich in diversity and encompasses numerous true-to-life situations that an urban environment has to offer. Within the dataset, over 5000 images measuring 1024 × 2048 pixels were included. Cameras mounted at the front of vehicles provided a single viewpoint for each image collection, which was suitable for conducting thorough analysis of urban sceneries. There exists a strong association between each image and precise pixel semantic annotations, which micro classifies the image at the level of single pixels into a total of thirty categories. Urban environments are graphically classified into several types including, but not restricted to, road, sidewalk, building, vehicle, pedestrian and traffic light. The dataset contains two types of annotation. Fine annotation is the annotation put at the individual pixels and is subsequently used for model training and evaluation. Coarse annotation, on the other hand, is less detailed but more generalized which can be useful in tasks like pretraining and semi-supervised learning. The dataset is already split into training, validation, and test sets maintaining uniformity and standardization for evaluation of semantic segmentation models. The technologies used in the manual collection as well as the annotation have been calibrated in a way such that the Cityscapes dataset provides

a valuable and diverse tool for facilitating research directed at urban scene interpretation applications in areas like self-driving cars, smart city development, and many more.

B. Data Preprocessing

Data Preprocessing comprises of segmentation and Zoom Out Field of View which serve the purpose of masking the input images and preparing them for 3D reconstruction. OneFormer which is a transformer based used for segmentation classifies regions of an image such as roads, vehicles etc. which is useful for object recognition and spatial analysis as it obtain the object boundaries with accuracy. As a result of adding padding, the Fold of view step increases the area that can be seen and reconstruct the image filling in missing parts and estimating depth enhancing the context. This guarantees that occluded or truncated structures, for instance, cars with only partial views, are constructed properly during the modeling which is an essential component of 3D reconstruction.

- Segmentation is the process in which the input image is separated into various regions or objects to show important features such as roads, buildings, vehicles, and pedestrians. This process is necessary for spatial understanding because it distinguishes the foreground and the background and delineates objects. Segmentation also provides bounding boxes for the objects in the image which in turn makes the subsequent tasks of object detection and spatial relationship easier. The model OneFormer is used in this experiment since it is specialized in producing high-quality segmentation masks of urban areas, which tend to be crowded and varied. Its accuracy makes sure that important parts of the scene are clearly indicated, which is very important for the later processes in the pipeline.
- The Zoom Out FoV step addresses the problem of occluded and partially visible objects in the scene by augmenting the input image area. It handles the problem of padding as well as zoom out to achieve a wider context for inpainting and depth estimation. The larger context guarantees that even those elements of objects which are occluded or truncated are taken into account so that reconstruction is done with all these elements. For example, a larger car which is only a part of the image due to constraints of the frame will cause the zoom out effect to allow the pipeline to infer and reconstruct the missing sections and thus the final output will contain all the sections and ensure geometric and visual continuity. This step is key in producing accurate 3D reconstruction without any visible seams.

C. Analysis and model selection

For generation of descriptions, two models were studied: BLIP (Bootstrapped Language-Image Pretraining) and LLaVA (Language and Vision Assistant). BLIP demonstrates outstanding performance on multi- modal tasks, such as image captioning as well as cross-modal retrieval and image description generation. Nevertheless, it functions as an independent system, which means that other procedures must be performed first to prepare the image for segmentation and spatial comprehension. On the contrary, LLaVA improves computational efficiency and complicates the pipeline by combining the generation of a description and the segmentation into one framework. Liking to describe and segment simultaneously makes it easy to deal with challenging scenarios like urban cities. For these reasons, LLaVA was picked precisely because of its dual functionalities which aid in smoothing the preprocessing workflow and the system performance as a whole. This variant makes sure to capture the intent in an accurate way while also being very optimal.

D. Experimental setup

The experimental design of the suggested framework incorporates a fusing of advanced tools and models concerning achieving high quality end product 3D reconstruction from a single view RGB image. The components of the pipeline were also fused sequentially throughout to ensure that a high task relevance concerning individual components within the framework was supplied. For the purpose of boosting the interpretive quality of input images, we used the vision-language models BLIP (Bootstrapped Language-Image Pretraining) and LLaVA (Large Language and Vision and Assistant). These models generated textual captions which enhanced the interpretability of the scenes. This step was important for throwing a shine on the assignment as it made it possible to augment the dataset with multimode information and advanced the object recognition and segmentation tasks. For segmentation, OneFormer model which is based on transformer architecture was applied since it is good for tasks that require both semantic segmentation and scene classification. The model successfully segmented the images into meaningful components such as road, buildings and vehicles which was required for further depth and geometric processing. The inpainting process was supported by generative models such as Stable Diffusion SD 3.5 and Runway ML to recover hidden or missing parts in images. Among the two, Stable Diffusion SD 3.5 stood out for its average Structural Similarity Index Measure (SSIM) value of 0.48 which showed it could give coherent and visually realistic results compared to others. When this stage was over, complete visual information that is high quality was used in 3D reconstruction process. In order to generate depth maps, MiDaS, Depth-anything v2 small, and Depth Pro models were evaluated. Generation of point clouds, which are required for 3D reconstruction, was done with high accuracy using Depth-Anything v2 Small. The models used here estimated geometry from a single image d,thus reducing the necessity for additional multi-view imagery and sensors. At last, the Open3D library was employed for creating and processing point clouds. This library offered a collection of filtering, downsampling, and 3D data reconstruction tools, helping further to mesh the point clouds. Its impressive performance with 3D data brought its usefulness to spatial analysis and visualization. This experimental setup shows a well-designed pipeline that incorporates cutting edge models and libraries to obtain de-

tailed and efficient 3D scenery reconstruction. Capitalizing on each component's strength, the such framework integrates both closely plausibility in geometric modelling as well as visual representation which broadens the scope of usage of such frameworks.

E. Reliability and validity

The reliability and validity of the framework is ensured through the use of strong models, standardized workflows, and quantitative evaluation metrics. The aforementioned is achieved by employing state-of-the-art patterns such as LLaVA, OneFormer, Stable Diffusion as well as Depth-anything v2 that are tested and proven to be accurate across a diverse range of datasets. This innovative construction, based on the use of a modular pipeline, allows the independent operation of every intra-system component while providing structural synchronism to the system for diversified input scenarios. Also, libraries such as Open3D typecast point cloud generation and mesh generation processes ensuring repeatability of results. The Structural Similarity Index Measure (SSIM) and point cloud accuracy are such metrics which help ascertain if the intermediate and final outputs have remained stable over time or at different instances. The validity of the framework is further reinforced by its ability to reconstruct single view RGB images of real world scenes. The semantic validity of the scene context is guaranteed by the description generated by the LLaVA in combination with mask ES which obtain meaningful object regions and boundaries. Furthermore, models such as MiDas and Depth-anything v2 can deliver accurate and complete representation of spatial relations without the need for multi view input. Integrating innovative inpainting models such as stable diffusion SD 3.5, which fills in the missing regions makes sure that there is visual and geometric consistency, or in other words, the final render is correct. Using Open3D to construct the final three-dimensional representations provides both the real life appearance and accuracy of geometry without compromising on the final aim of VR which is providing an exact and detailed depiction of real world environments. Performance measures of geometrical accuracy such as point cloud density and view quality like SSIM serves to norm global to the robustness and reliability of the framework serving for self-navigating cars, city development and VR.

F. Evaluation metrics

The Structural Similarity Index Measure (SSIM) Mean Scores for Runway ML and Stable Diffusion 3.5 are compared using various test images in the Table 1. The process of evaluating reconstructed images resemblance to the original images is done with the help of SSIM, which is commonly used. When the SSIM values tend to be higher, the images tend to be more structurally and visually consistent. This evaluation shows how the two models compare in terms of their strengths and functional capacity within various models.
Table 1: Comparison of Inpainting Models

All the tested images show that the model performance of Stable Diffusion 3.5 is significantly better than Runway ML with higher SSIM scores. The average SSIM for Stable Diffusion 3.5 is 0.482800 while for Runway ML it's at 0.445595 showing an improvement of approximately 0.037206 which translates to roughly a fifth of what it was before. That suggests that among the two inpainted image generators, Stable Diffusion 3.5 creates more coherent and accurate visual representations in terms of structure particularly in comparison to Runway ML. With a range in SSIM scores from 0.020645 to 0.070744, we can conclude that this gap in performance varies based on how intricate or simple each image appears. The SSIM difference for simple images like stuttgart000029000019leftImg8bit is calculated to be around 0.011498 which indicates the two models are performing similarly. However, for more complex SSIM images such as, stuttgart000033000019leftImg8bit, the difference becomes more significant at 0.070744 which indicates the modeling capabilities of capable Stable Diffusion 3.5. This demonstrates the efficiency of Stable Diffusion 3.5 through the ability to better inpaint structural details in an image. The Stable Diffusion 3.5 has shown reliable inpainting results by achieving consistently high SSIM scores across all images, which indicates its effectiveness in producing good quality additions. Such reliability is pertinent for tasks that necessitate meticulous visual and geometric recovery like 3D reconstruction, self-driving vehicles to autonomous robots, or construction of virtual spaces. The further ability of Stable Diffusion 3.5 to modify with respect to the image context makes it more flexible for inpainting tasks. As the assessment indicates, inpainting is better carried out with Stable Diffusion 3.5 than with Runway ML. The higher SSIM scores that they achieved are indicative that the inpainting wthat 'graphics does or looks like' is structurally correct. As such, such tools are useful for advanced inpainting. The continuous enhancements across different settings demonstrate its resilience and flexibility, corroborating its utilization in complex computer vision tasks.

IV. Experimental results

The given images depict the result of inpainting by showing a "Before Image" with missing or occluded parts and an "After Inpainting Image" which fills these gaps. Relevant parts of the architectural, vehicular centre, road and other urban street objects in the older image are either occluded partially or fully making the entire scene incomplete. Such details prohibit the image from being used in further applications for instance depth estimation or 3D reconstruction and other purposes. For the "After Inpainting Image", the inpainting model utilizes the information captured in the context of the visible parts of the image to fill in the missing details and make the image more coherent, thereby reconstructing architectural details, textures of road, and other surrounding objects with high accuracy. The inpainting model utilizes best pixel prediction strategies and takes into consideration the preexisting parts of the image providing predictions for the non visible details of the image. The spatial relationship and the perspective in

which the fragments and constructs are constructed is retained, ensuring that the buildings and cobblestone streets constructed appear integrated into the original picture. Furthermore, the advanced visual representation that is presented ensures that the image decorum fits other preprocessing tasks like segmentation, depth maps as well as 3d modeling, thus, making them much more efficient. The findings show the capability of the inpainting model in recovering very complicated scenarios, specifically in urban settings comprising numerous interacting objects, and with complex buildings. The enhancement has proven the model to be trustworthy which produces visually credible and geometrically accurate images, thus provides for suitable implementation in autonomous navigation, urban modeling and smart city building.

Inpaint Results The images in Fig 4 are before inpainting whereas Fig 5 is reconstructed into after Inpainting Images with certain regions filled in. The "Before image" is said to be made up of an urban street scene in which some of the architectural elements like the road, houses, cars are either not fully exposed or have been covered, which reduces the completeness of the image. These details that are supposed to be in the image are crucial for its deployment in later tasks including the estimation of depth and in 3D reconstruction, thus making them mostly useless.

Fig. 3. Original image

In Fig 4 it is shown that this process works well by filling up occulsions and missing regions thereby, reconstructing architectural details, road textures, and surrounding elements with high visual consistency. The inpainting model utilizes contextual information from the visible parts of the image to predict and synthesize the missing details thus leading to a natural and coherent scene. Reconstructed buildings, cobblestone road as well as urban elements merge seamlessly with the original content preserving spatial alignment and perspective. This improved visual representation makes the picture more appropriate for subsequent processing associated with segmentation, depth mapping or 3D modeling.

Fig. 4. Image After In-paint

The findings show the effectiveness of the inpainting model in reconstructing a complex urban scene which is characterized by having multi-interacting objects and detailed urban structures. The advancement emphasizes the trustworthiness of the model to produce realistic and geometrically coherent outputs making it a useful method for tasks like autonomous driving, urban modeling, and smart cities building.

- MiDaS: It is Efficient but Limited in Detail. MiDaS creates depth maps which include camera depth presets of definite angles that encompass all of the essential parts of a scene such as the transitions of due depth shifts of the background as well as objects in the foreground. From this information, it may be pointed out that its corresponding Point Clouds lack detail and the finer geometric details are left up reflecting the already mentioned limitations. Hence, MiDaS is speed efficient and can be used in general-purpose depth estimation tasks, however, due to having inability to capture the intricate geometry of object contours, 3D reconstruction for such objects which requires high accuracy is not optimal for MiDaS.
- Depth Pro: It gives a Balanced Performance. It consistently produces significantly better quality depth maps with sharp edges, better contrast, and accurately identifying roads, poles, and buildings around the scene. The formed point cloud bears more geometric accuracy than that produced by MiDaS as "surfaces and edges are better

defined". This makes Depth Pro an adequate solution for applications with limited computational power and a requirement of minimal accuracy. However, such conditions lead to subpar performance when the scene is highly detailed with a large number of objects or slight depth complexity.

- DepthAnything V2: It produces Superior Detail and Precision in depth maps. It is the best for the most detailed results and it captures intricate information about the scene better than any other system. For example, its depth maps are highly accurate, and the resultant point cloud is thick and nicely organized to represent complex geometries as well as subtle depth variations accurately. Such high level of accuracy makes DepthAnything V2 a good fit for producing 3D reconstructions at a premium quality especially in areas such as independent vehicle navigation, town planning and virtual environment development. Nevertheless, this may limit its applicability under resource constraints due to increased computational loads.

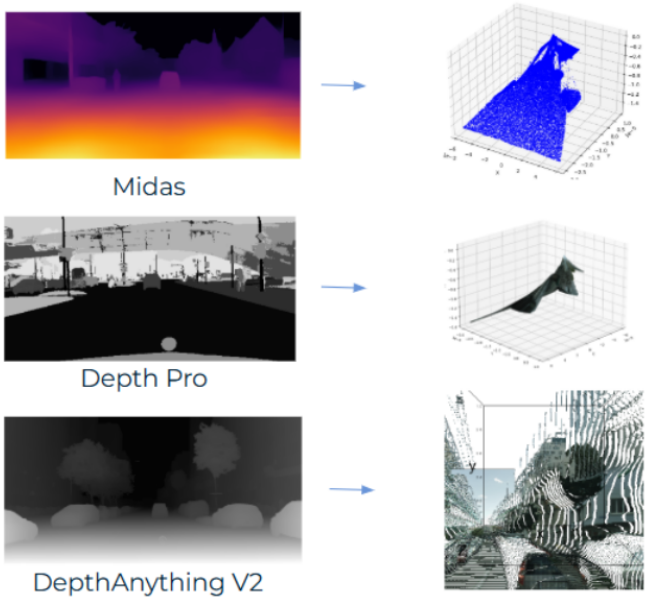

Fig. 5. Performance Comparison of Depth Estimation Models

To sum up, the selection of depth estimation model is in direct response to the needs of the application. For approximate depth estimation of the federated embedding with the least amount of calculation difficulty, MiDaS suffices. Depth Pro addresses middle range tasks dwelling on its coordination between efficiency and accuracy. At the same time, DepthAnything V2 gives the greatest accuracy but only for digital, 3D models taking intricate, detailed images through 3D reconstruction - at a steep, high, computational cost. This kind of analysis allows one to realize the meaning and purpose of models in relation to the balance between performance, accuracy and computational efficiency. A pipeline for 3D reconstruction from a single-view image is illustrated by the visual starting with an original RGB image. The progress commences with segmentation, which identifies areas like roads, vehicles, and buildings and thus gives them semantic context. Next, the segmented image becomes larger via zoom-out and missing parts are filled in with inpainting to end up with a whole image of the scene that makes sense visually. The next step generates depth map where each pixel gets its depth value disclosing the spatial geometry and interconnections within the scene. In depth analysis, along with the pre-processed image, visual information is mapped into the depth map, consolidating geometric and visual information. From this depth map enriched, a point cloud is created, which patches the 3D scene comprising millions of data points. In the hypothetical case, preprocessing discover around 8 million data points, and 500,000 of these points are selected to render an accurate and superb quality 3D point cloud. This pipeline provides high 3D reconstruction accuracy suitable for urban modeling, autonomous navigation and virtual environment applications.

V. Conclusion

In this work, an effective transformer-based framework for single-view 3D reconstruction is presented, showcasing its capacity to tackle important issues including structural coherence, occlusion, and depth estimation. The suggested pipeline effectively bridges the gap between 2D and 3D representations by combining cutting-edge vision-language models (BLIP, LLaVA), segmentation tools (OneFormer), generative inpainting models (Stable Diffusion SD 3.5), and sophisticated depth estimation techniques (Depth-Anything v2). The framework's effectiveness is confirmed by evaluation on the Cityscapes dataset, which shows great structural fidelity with an SSIM score of 0.48. This makes the framework a promising solution for a variety of applications in urban planning, autonomous cars, and augmented or virtual reality environments.

The pipeline's modular architecture promotes flexibility and adaptation, facilitating the smooth integration of sophisticated models and methodologies while preserving computing efficiency. Additionally, the framework's scalability and accessibility for both academic and industrial research are highlighted by the usage of open-source technologies like Open3D.

This work lays the groundwork for future developments in 3D scene reconstruction while also demonstrating the revolutionary potential of transformer-based systems in computer vision. Enhancing accuracy and applicability in intricate, resource-constrained contexts may be possible by extending the framework to include multimodal data like LiDAR, thermal imaging, or multiview sampling. This study opens the door to novel solutions in autonomous systems, immersive virtual worlds, and smart cities by pushing the limits of single-view reconstruction.

VI. Future Work

In terms of future improvement options relevant to the framework model, a number of factors have been outlined. One such factor relates to the integration of more sensory information such as LiDAR or infrared imaging into the

framework model in case light levels are too low or the environment too complicated in a way that RGB images are not sufficient. Such a multimodal approach would allow for expansion of the functionality of the system to cover a wider variety of tasks and make it more resilient. Furthermore, the framework could utilize multiview consistency sampling, even in the context of a single-view approach. By applying methods that maintain spatial coherence of recreated scenes, it would be possible to improve the quality of 3D models, closing the gap between single-view images and true multi-view models. These improvements would broaden the range of prospective applications of the framework, preventing stagnation of the progress made in 3D scene understanding and reconstruction.

REFERENCES

[1] Y. Liu, C. Lin, Z. Zeng, and T. Komura, "SyncDreamer: Generating Multiview-Consistent Images from a Single-View Image," in Proceedings of ICLR 2024.

[2] H. Wang, Y. Liu, and B. Yang, "VistaDream: Sampling Multiview Consistent Images for Single-View Scene Reconstruction," arXiv preprint arXiv:2410.16892, 2024.

[3] H. Li, Z. Liu, and M. Zhang, "LLaVA: Large Language and Vision Alignment for Image Captioning," arXiv preprint arXiv:5336.02238, 2022.

[4] Li, Y. Lin, and X. Zhang, "BLIP: Bootstrapping Language-Image Pretraining for Unified Vision-Language Understanding," arXiv preprint arXiv:2201.12086v2, 2022.

[5] Z. Shi, J. Chen, and Y. Li, "OneFormer: A Universal Image Segmentation Framework," arXiv preprint arXiv:2211.06220v2, 2022.

[6] A. Ramesh, M. Pavlov, and G. Kim, "Stable Diffusion: Score-Based Generative Modeling for Image Synthesis," arXiv preprint arXiv:2403.03206v1, 2024.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv preprint arXiv:2112.10752v2, 2022.

[8] Reiner Birkl, Diana Wofk, and Matthias Müller, "MiDaS v3.1 – A Model Zoo for Robust Monocular Relative Depth Estimation," arXiv preprint arXiv:2307.14460, 2023.

[9] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun, "Depth Pro: Sharp Monocular Metric Depth in Less Than a Second," arXiv preprint arXiv:2410.02073, 2024.

[10] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao, "Depth Anything V2," arXiv preprint arXiv:2406.09414, 2024.

[11] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A Modern Library for 3D Data Processing," arXiv preprint arXiv:1801.09847, 2018.

[12] Fu, K., Peng, J., He, Q. et al. Single image 3D object reconstruction based on deep learning: A review. Multimed Tools Appl 80, 463–498 (2021). https://doi.org/10.1007/s11042-020-09722-8

[13] Han, X., Laga, H., and Bennamoun, M. (2019). Image-based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era. ArXiv. https://doi.org/10.1109/TPAMI.2019.2954885

[14] Henderson, P., and Ferrari, V. (2019). Learning single-image 3D reconstruction by generative modeling of shape, pose, and shading.. ArXiv. https://arxiv.org/abs/1901.06447

[15] Shi, Z., Meng, Z., Xing, Y., Ma, Y., and Wattenhofer, R. (2021). 3D-RETR: End-to-End Single and Multi-View 3D Reconstruction with Transformers. ArXiv. https://arxiv.org/abs/2110.08861 https://doi.org/10.1016/j.renene.2016.12.095

[16] Wang, H., Liu, Y., Liu, Z., Wang, W., Dong, Z., and Yang, B. (2024). VistaDream: Sampling multiview consistent images for single-view scene reconstruction. ArXiv. https://arxiv.org/abs/2410.16892

[17] Zou, Z., Yu, Z., Guo, Y., Li, Y., Liang, D., Cao, Y., and Zhang, S. (2023). Triplane Meets Gaussian Splatting: Fast and Generalizable Single-View 3D Reconstruction with Transformers. ArXiv. https://arxiv.org/abs/2312.09147