Building the world with a drone: efficient creation of aerial scenes through localization, recognition and virtual view synthesis

Costea Dragos

1 Introduction

This section is an overview of the thesis, dealing the potential of vision-based navigation and scene understanding, focusing on aerial imagery analysis and reconstruction. The main questions addressed are:

- Can we navigate an environment using only camera input?
- How well do we need to understand the environment for effective navigation?
- How can we leverage and improve existing map data?
- Can all basic processing be performed on a drone?

The research is divided into four main areas: Localization, Mapping, Scene Reconstruction, and Interactive Scenes.

1.1 Localization

Localization is crucial for autonomous navigation, especially in urban environments where GPS signals may be unreliable. Vision-based localization offers a robust alternative, drawing inspiration from human visual capabilities. The challenge lies in developing efficient algorithms that can run on limited compute budgets typical of drone hardware.

Recent work in aerial image localization has shown promising results. For instance, [1] proposed a method for ground-to-aerial geolocalization by learning a shared embedding space for both ground and aerial images. [2] extended this idea to wide-area image geolocalization using aerial reference imagery.

1.2 Mapping

Building on localization, robust road mapping is essential for navigation and can provide a foundation for using intersections as localization landmarks. The goal is to bridge the gap between existing map data and real-time visual information.

Early approaches to road detection in aerial images relied on manually designed features [3–5]. Deep learning techniques have led to significant improvements, with [6] and [7] among the first to apply convolutional neural networks (CNNs) to this task.

More recent approaches have focused on leveraging context and multi-scale information. [8] proposed a dual-stream network that processes both local and global context, showing particular effectiveness in handling complex urban scenes.

1.3 Scene Reconstruction, Navigation, and Representation

This area focuses on creating accurate world replicas for advanced simulations, heritage preservation, and enhanced user experiences. It explores the synergy between geometric and analytical methods for novel view synthesis and investigates efficient representations for low-cost, vision-focused navigation systems.

Depth estimation from monocular images has seen significant progress, driven by advances in deep learning. [9] demonstrated the feasibility of training CNNs to predict depth from single images. Unsupervised and self-supervised learning techniques, such as those introduced by [10] and [11], have gained popularity as they alleviate the need for expensive ground truth depth data.

The field of 3D reconstruction has seen a paradigm shift with the introduction of Neural Radiance Fields (NeRF) by [12]. NeRF represents scenes as continuous volumetric functions and has demonstrated impressive results in novel view synthesis. Subsequent work, such as Instant-NGP [13], has focused on improving efficiency and scalability.

For large-scale aerial reconstruction, traditional photogrammetry techniques remain widely used. Structure-from-Motion (SfM) pipelines, such as COLMAP [14], can reconstruct 3D scenes from large collections of unordered images. However, these methods often struggle with the scale and complexity of city-scale reconstructions.

1.4 Interactive Scenes

The final area addresses the challenges of generating views for imperfect replicas and integrating virtual characters into reconstructed environments. It also explores the potential of generative art in human-machine interaction, particularly in conveying emotions.

In the context of human-AI interaction, there has been growing interest in generating appropriate emotional responses from virtual agents. [15] have explored the use of computational models of emotion to drive the behavior of virtual characters. [16] has investigated data-driven approaches to generate realistic facial expressions for virtual agents in real-time.

1.5 Motivation and Applications

The motivations behind this research include:

- Developing robust, vision-based localization systems for airborne devices to enhance safety in urban areas.
- Demonstrating the feasibility of vision-only navigation, building on successes in niche areas like drone racing [17] and simulated environments [18].
- Enabling automatic, large-scale map updates through the combination of precise localization and visual recognition.
- Improving 3D reconstruction quality by exploiting complementary information from classical and neural methods.
- Creating plausible novel views to fill information gaps in reconstructed scenes.
- Designing efficient algorithms suitable for embedded use, ensuring autonomous agents can operate safely even in the event of network failures.

1.6 Related Work

The field of aerial image analysis has evolved from early approaches using manually designed features [3–5, 19, 20] to modern deep learning techniques. Significant advancements have been made in semantic segmentation, with methods like those proposed by [8] and [21] improving accuracy in complex urban scenes.

Geolocalization research has progressed from traditional feature-matching methods [22, 23] to deep learning approaches that learn robust representations [1, 2]. In the context of UAV navigation, vision-based localization has been explored as an alternative or complement to GPS [24].

Depth estimation and 3D reconstruction have seen significant advancements, from early learning-based approaches [9] to more recent unsupervised techniques [10, 11]. The introduction of Neural Radiance Fields (NeRF) [12] has revolutionized novel view synthesis, with subsequent work focusing on improving efficiency and scalability [13].

Safe landing area estimation for UAVs has evolved from methods based on hand-crafted features [25] to deep learning approaches [26]. Recent work has explored the use of synthetic data to address the limited availability of labeled training data [27].

In the field of human-AI interaction, research has progressed from traditional approaches to emotion recognition [28] to deep learning methods [29]. Work on generating appropriate emotional responses from virtual agents [15] and real-time facial expression generation [16] has opened new avenues for natural and intuitive interfaces.

1.7 Conclusion

This research stands at the intersection of computer vision, robotics, and human-computer interaction. It builds upon existing foundations while introducing novel techniques to advance the state-of-the-art across these interconnected domains. The work has potential implications for autonomous navigation, urban planning, 3D modeling, and the development of more intuitive AI systems.

By addressing these challenges, the research aims to contribute to the development of more robust, efficient, and intelligent autonomous systems capable of understanding and interacting with complex, real-world environments. The interdisciplinary nature of the work highlights the potential for cross-pollination of ideas between different subfields of computer vision and robotics.

2 Localization from roads and intersections in aerial images

2.1 Introduction

Aerial image analysis has important applications in automated mapping, urban planning, environment monitoring and disaster relief. This chapter addresses the task of automatic geolocalization of aerial images from recognition and matching of roads and intersections. We propose a novel pipeline for geolocalization, from road and intersection detection to identifying the geographic region by matching detected intersections to manually labeled ones from OpenStreetMap (OSM). This is followed by geometric alignment between detected roads and OSM annotations. We test on a dataset of aerial images from two European cities and use OSM for ground truth road annotations. Experiments show accurate localization when training on one city and testing on the other, even with relatively poor quality aerial images. We also demonstrate that alignment between detected roads and OSM annotations can improve road detection quality.

2.2 Related Work

Road detection in aerial imagery has been addressed using manually designed features [3, 4] and more recently convolutional neural networks [6, 7]. Some methods attempt to correct misaligned road vectors by aligning them to aerial images [30]. Geolocalization for UAVs using sparse manually designed features has been proposed [24]. Other approaches fuse camera input with GPS and IMU data [31, 32]. Geolocalizing ground images using aerial image pairs has also been explored [1, 2].

2.3 Approach

Our method has several stages:

Road pixel-wise classification using a dual stream local-global CNN [33]. Intersection detection based on the detected roads. Matching detected intersections to a stored dataset of OSM intersections using learned descriptors. Geometric alignment for improved localization and road detection enhancement.

We use intersections as anchors for localization as they are sparse, computationally efficient, and tend to have unique surrounding road patterns useful for recognition.

2.3.1 Road and Intersection Detection

For road detection, we use a state-of-the-art dual stream local-global CNN [33] that combines local appearance and larger contextual information. For intersection detection, we train an adjusted AlexNet that takes as input the RGB image and estimated road map.

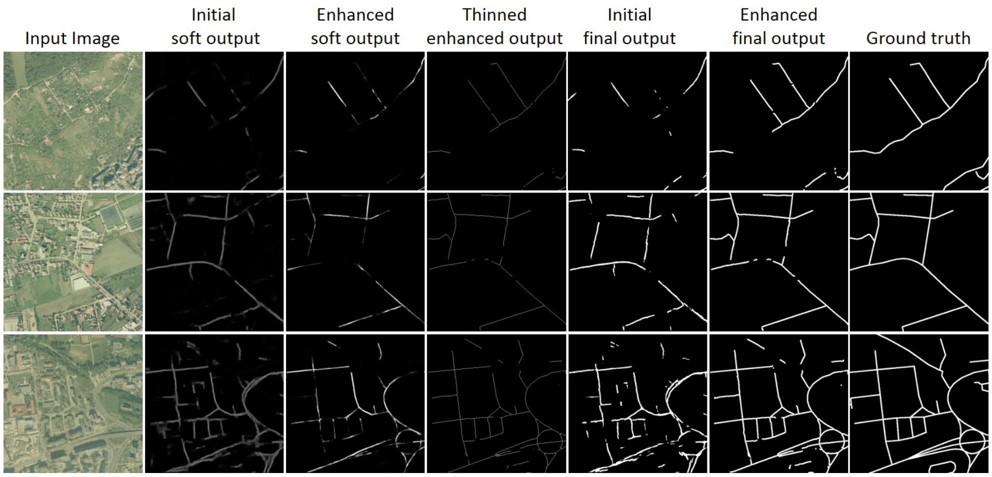


Figure 1: Enhancing road detection by region recognition and geometric alignment to OSM roads. Our procedure improves detected road maps and could correct OSM labels.

Table 1: Localization errors before and after alignment (in meters)

Method	Before		After	
	Mean	Median	Mean	Median
1NN pix2pix	470.50	53.90	57.97	1.60
1NN MSMT Stage 1	34.11	31.65	454.79	1.40
MSMT with LocDecoder-R-2	88.92	53.90	57.97	1.60
MSMT with LocDecoder-S-128	**9.03**	**6.85**	**1.89**	**0.75**
MSMT with LocCombined	26.93	7.27	18.42	0.78

2.3.2 Intersection Matching and Localization

We represent each intersection by a learned descriptor such that identical intersections from detected and OSM roads have similar descriptors. We fine-tune the intersection detection network to adjust distances in descriptor space and improve matching performance. Localization is refined by geometric alignment between estimated roads and OSM roads in regions centered at matched intersections. We use a bipartite graph matching approach to find correspondences between detected and OSM intersections.

2.3.3 Geometric Alignment and Road Enhancement

We use the Iterative Closest Point algorithm to align the road segmentation with OSM roads at the predicted location. We developed a simplified version that only estimates translation, assuming images are aligned with cardinal points. To enhance road detection, we apply soft dilation on the estimated road map, multiply it with the aligned OSM map, smooth with a Gaussian filter, and thin using non-maximum suppression.

2.4 Experiments

We collected aerial images of two European cities (A and B) aligned with OSM road maps. City A has 4027 512x512 pixel images used for training, while city B has 3177 images used for testing. The spatial resolution is 1m/pixel. For road and intersection detection, we achieve an F-measure of 82.22% on the European Roads Dataset [34] and 99.88% on our localization dataset with 3-pixel relaxation. For localization, 96.84% of test locations have an error ¡20m without alignment. After alignment, 94.56% are within 2.5m and 97.58% within 5m of ground truth, comparable to commercial GPS accuracy [35].

2.5 Conclusions

We presented a complete system for geolocalization from aerial images without GPS. Our pipeline includes efficient methods for road and intersection detection, intersection recognition with geometric alignment for accurate localization, and road detection enhancement. The approach could be used as a GPS alternative or in conjunction with GPS for applications requiring offline or real-time processing. Future work could focus on improving detection speed, expanding the search space to multiple cities, and adapting the pipeline for nighttime use.

3 Detecting roads and buildings in aerial images

3.1 Introduction

Recognizing roads and intersections in aerial images is a challenging problem in computer vision with applications in UAV localization and navigation. While recent deep learning approaches have improved pixel-level segmentation, we argue roads should be recognized at the higher semantic level of road graphs. We present a two-stage method: 1) Detect roads and intersections with a novel dual-hop generative adversarial network (DH-GAN) for pixel-level segmentation. 2) Find the best covering road graph using smoothing-based optimization (SBO). We also present a multi-stage improvement of this pipeline.

3.2 Road Detection with DH-GAN and SBO

Our DH-GAN architecture consists of two conditional GANs: the first generates road segmentations, while the second generates intersections using both the original RGB input and the road segmentation from the first GAN (Figure 2). The full architecture is trained end-to-end. We represent roadmaps as graphs $G = (V, E)$, where V are nodes with positions (x_i, y_i) and E are edges with associated line segments. We define edge cost $c(i, j)$ as the average distance to the nearest road point in the segmentation, and overall graph score $S(V)$ as the intersection over union between the dilated graph and pixelwise map. Our SBO approach optimizes node locations to maximize $S(V)$, starting from intersections found by DH-GAN and iteratively adding midpoints. We also apply a greedy sampling baseline for comparison. We evaluate on the European Road Dataset [33], with 200 training, 20 validation, and 50 test images. Tables 2-3 show our results compared to baselines. DH-GAN outperforms other methods in pixel-level accuracy. The graph-based approaches (DH-GAN+Greedy and DH-GAN+SBO) trade some accuracy for significant storage savings (Table 4). Qualitative results are shown in Figure 3.

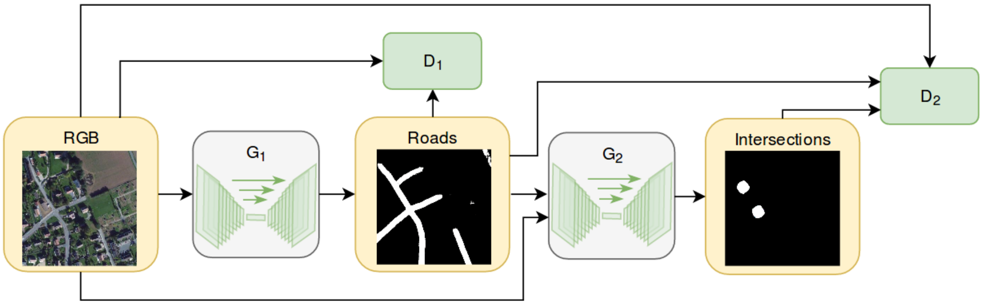


Figure 2: Proposed DH-GAN architecture.

3.3 Multi-Stage Ensemble Approach

We propose a three-stage method for road extraction:

1. Train multiple U-net-like networks with different dilation rates for road and intersection segmentation. 2. Combine partial predictions with RGB input in another network for improved segmentation. Generate road vectors using SBO. 3. Add missing links using both segmentation and road vectors.

Method	F-measure
GAN [36]	77.70%
LG-Seg-ResNet-IL [37]	81.06%
U-net [38]	79.79%
DH-GAN	**84.05%**
DH-GAN + Greedy	80.81%
DH-GAN + SBO	81.74%

Table 2: Road detection results. Higher is better.

Labeling Type	Method	F-measure
OSM	GAN [36]	54.89%
	DH-GAN	**63.01%**
	DH-GAN + Greedy	31.81%
	DH-GAN + SBO	59.79%
Independent	GAN [36]	64.42%
	DH-GAN	82.65%
	DH-GAN + Greedy	43.42%
	DH-GAN + SBO	**86.00%**

Table 3: Intersection detection results. Higher is better.

Method	Vertices	Edges
LG-Seg-ResNet-IL [37]	782*	-
DH-GAN	1345*	-
DH-GAN + Greedy	23	21
DH-GAN + SBO	**19**	**17**
OSM (ground truth)	29	31

Table 4: Average storage cost. Lower is better. *Vertices obtained by thinning roads to single pixel width.

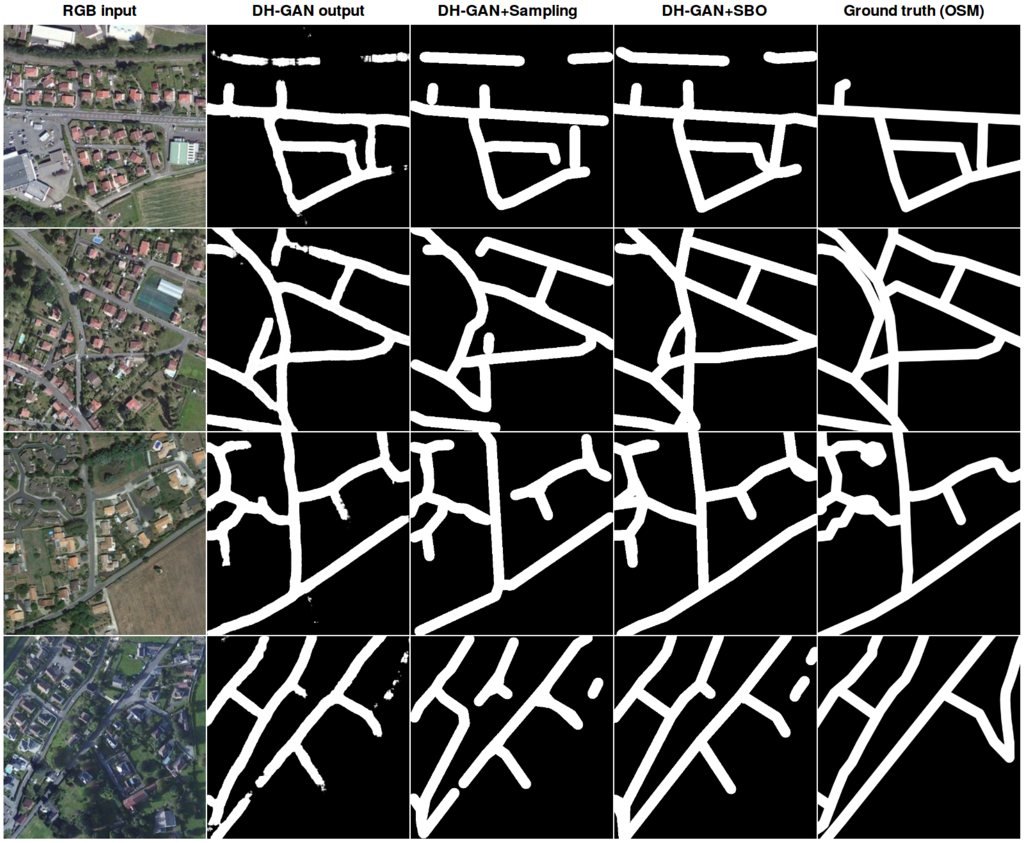


Figure 3: Qualitative results for road detection and map generation using graphs.

Table 5: Roads segmentation results on our training/validation split.

Model	Our Training		Our Validation	
	IoU	F1	IoU	F1
Max dilation 32	0.6432	0.7824	0.6483	0.7883
Max dilation 48	0.6577	0.7913	0.6601	0.7957
Max dilation 64	**0.6591**	**0.7919**	**0.6640**	**0.7966**

We use the DeepGlobe Dataset [39] with 6226 training, 1243 validation, and 1101 test images. Our U-net variants use chained dilated convolutions with different maximum dilation rates (32, 48, 64). We also train a network on constant-width roads. Results are shown in Tables 5-**??**.

We found that training on thicker roads improved performance (Table 6).

3.4 Conclusions and Future Work

We presented two approaches for road detection in aerial imagery:

1. DH-GAN with SBO for efficient road graph extraction, combining deep learning and graph optimization.
2. A multi-stage ensemble approach using multiple dilation rates and road vector refinement.

Both methods show improvements over baselines. Future work includes improving existing maps, multi-level representations combining ground and aerial views, and addressing road width variations.

Table 6: Road thickness study using Max dilation 32 model.

		IoU Our Training	IoU Our Validation
Same width	Thin (≈ 4m)	0.6282	0.6123
	Thick (2x thin)	0.6918	0.6751
Variable width	Thin (original)	0.6432	0.6483
	Thick (2x thin)	0.7254	0.6889

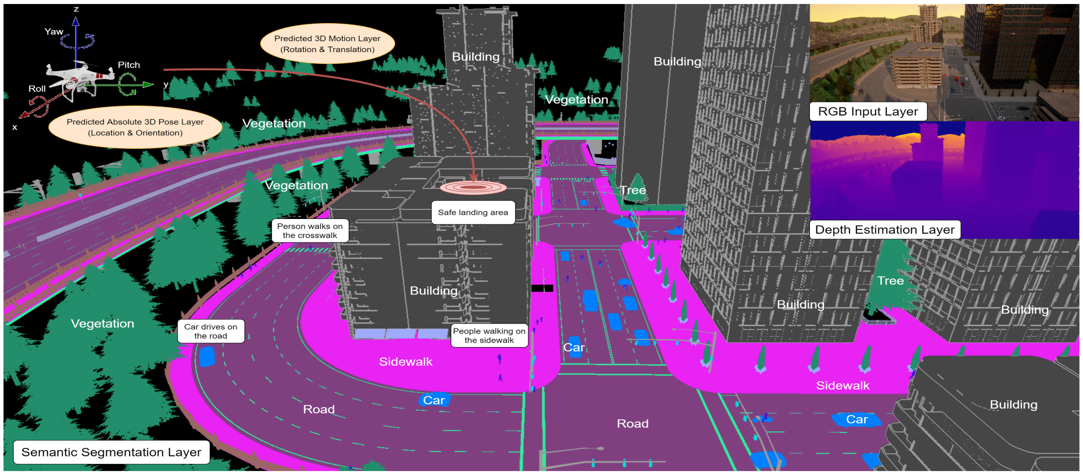


Figure 4: NGC can put together different interpretations of the dynamic scene, such as 3D structure, pose, motion, semantic segmentation of objects and activities in different regions of space and time, into a unified neural graph, in which multiple paths reaching a given node become teacher through consensual agreements to any single edge net reaching the same node. Trained in this self-supervised manner, NGC can reach robust unsupervised learning in the face of unlabeled data. The scene in the figure is taken from a virtual environment used to collect data for our experiments.

4 Towards a complete understanding of the world with a drone

We aim to better understand the world from multiple representations. First, we present SafeUAV, a safe landing solutions that leverages depth and surface normals to output safe landing regions. We then present Neural Graph Consensus, a self-supervised algorithm for improving scene understanding based on a diverse set of representations, as shown in Figure 4.

4.1 Safe landing for UAVs

We propose SafeUAV-Net, an embeddable system based on deep convolutional networks for depth and safe landing area estimation using only RGB input. We produce a synthetic dataset and train on it, showing compelling performance on real drone footage.

4.1.1 SafeUAV-Net for depth and plane orientation estimation

We aim to predict depth and classify plane orientation into three classes: horizontal, vertical and other. Our tasks are related to semantic segmentation as we predict a categorical value for each pixel. We use a variant of the U-Net model proposed by [40] for aerial image segmentation. We developed two variants: SafeUAV-Net-Large runs at 35 FPS on Nvidia's Jetson TX2, while SafeUAV-Net-Small runs at 130 FPS. The detailed architectures are described in Figure 5.

4.1.2 Dataset

We construct our virtual dataset using Google Earth [41] 3D reconstructions. The dataset consists of 11,907 samples in an 80

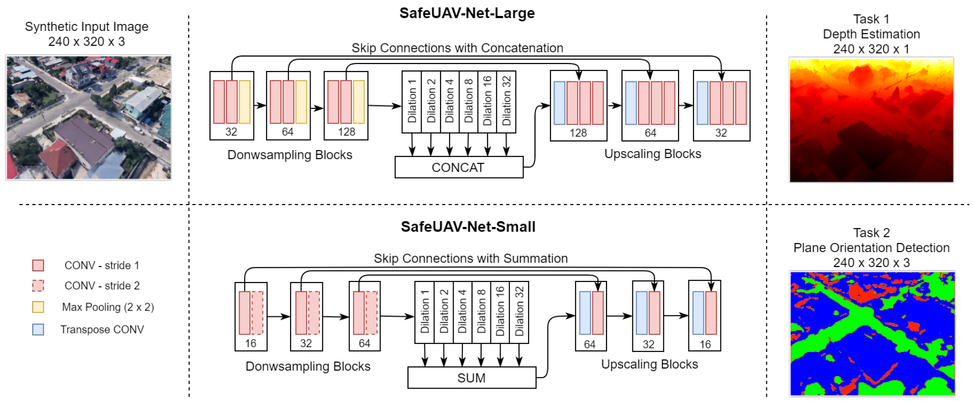


Figure 5: Our proposed SafeUAV-Nets for both on-board and off-board processing, trained for depth estimation and plane orientation prediction.

Table 7: HVO prediction results for SafeUAV-Net-Large and SafeUAV-Net-Small trained on full dataset.

Model	Input	Accuracy	Precision	Recall	mIoU
U-net [38]	RGB	0.729	0.560	0.505	0.356
DeepLabv3+ [42]	RGB	0.840	0.753	0.739	0.597
Small	RGB	0.823	0.728	0.693	0.551
Large	RGB	**0.846**	**0.761**	**0.748**	**0.607**

4.1.3 Experiments

We report qualitative and quantitative results on depth estimation and HVO segmentation on all four regions from our dataset. Tables 7 and 8 show the results. Our experiments on unseen synthetic test cases show that our system is numerically accurate while being fast on an embedded GPU. We believe the use of our approach on commercial drones could improve flight safety in urban or suburban areas at high speeds and complement on-board sensors.

4.2 Semi-Supervised Learning for Multi-Task Scene Understanding using Neural Graph Consensus

We propose Neural Graph Consensus (NGC), a novel model for semi-supervised learning of multiple scene interpretations. NGC connects multiple deep networks into a large neural graph, where each node represents a different interpretation of the scene (e.g., depth, semantic segmentation, pose). The

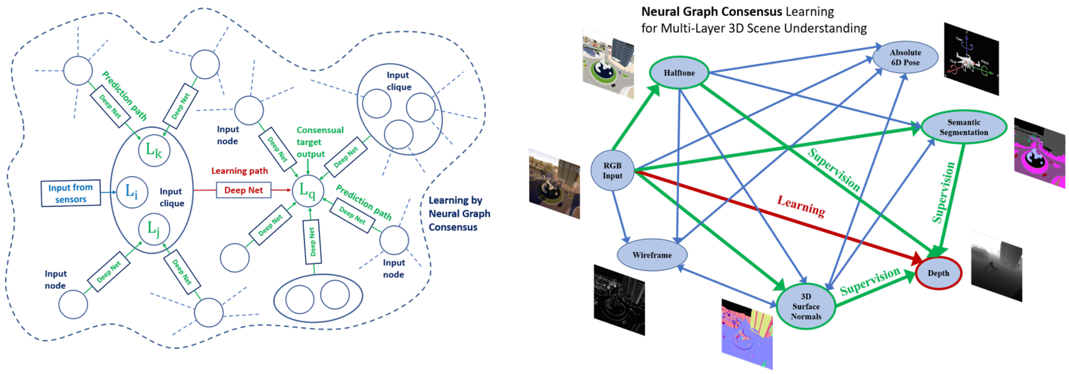
Table 8: Results on depth estimation for SafeUAV-Net-Large and SafeUAV-Net-Small trained on full dataset. Errors are expressed in meters.

Model	Input	RMSE	Meters
U-net [38]	RGB	0.041	9.63
DeepLabv3+ [42]	RGB	0.034	8.49
Small	RGB	0.031	7.22
Large	RGB	**0.026**	**6.09**

		Iteration 0	Iteration 1		Iteration 2	
Representation	Evaluation Metric	EdgeNet	NGC	Distil. EdgeNet	NGC	Distil. EdgeNet
Depth	L1 (meters)	4.9844	3.4867	4.2802	**3.2994**	**3.9508**
Surface Normals (C)	L1 (degrees)	8.4862	7.7914	8.2891	**7.4503**	**7.6773**
Surface Normals (W)	L1 (degrees)	11.8859	8.8248	10.7500	**8.5282**	**8.6714**
Semantic Segmentation	mIOU	0.4840	0.4978	0.4980	**0.5258**	**0.5159**
Wireframe	Accuracy	0.9617	0.9655	0.9654	**0.9661**	**0.9655**
Position	L2 (meters)	25.7597	15.5383	20.0204	**12.0764**	**15.5599**
Orientation	L1 (degrees)	3.8439	2.5001	3.3961	**2.2088**	**3.0005**

Table 9: Results for our proposed ensemble NGC and distilled EdgeNets on 6 representations, over 2 iterations of unsupervised learning.

edges between nodes are deep networks that transform one representation into another. The NGC model is illustrated in Figure 6.



A. General Conceptual NGC Model

B. Actual NGC model implemented

Figure 6: Neural Graph Consensus (NGC) model architecture

4.3 Neural Graph Consensus Model

Each node i in the graph has an associated layer L_i, encoding a specific view or interpretation of the space-time world. Layers at nodes can be predicted from other layers by deep nets forming edges in the graph. During unsupervised learning, each net becomes a student of the NGC graph and is trained by the mutual consensus from contextual pathways reaching the same output node. NGC becomes a self-supervised system where agreement is the ultimate teacher for unlabeled data.

4.3.1 Experimental Analysis

We capture a large dataset using a customized CARLA simulator [43], where a drone flies above a city predicting scene depth, 3D surface normals, absolute 6D pose, scene wireframe, and semantic segmentation from a single image. We developed a general NGC framework on top of PyTorch [44]. For EdgeNets we used Map2Map and Map2Vector architectures, each with about 1.1M trainable parameters. The total NGC model has 27 edge nets, totaling about 30M parameters. Table 9 shows results for our proposed ensemble NGC and distilled EdgeNets on 6 representations over 2 iterations of unsupervised learning. We compared NGC with state-of-the-art multi-task learning methods NDDR [45] and MTL-NAS [46], and semi-supervised learning method CCT [47]. Tables 10 and 11 show the results.

Figure 7 visualizes the improvements achieved by NGC over baseline single-task models for various scene understanding tasks.

Task	Metric	EdgeNet(iter 0)	NGC	NDDR*(no pretrain)	NDDR*(pretrain)	NDDR	MTL-NAS
Semantic Segm.	mIOU	0.484	**0.498**	0.141	0.343	0.315	0.368
	Acc.	90.017	**91.816**	48.7	84.2	86.9	87.8
Normals (C)	Err (deg.)	8.4862	7.7914	9.820	7.727	6.801	**6.533**

Table 10: Multi-task learning results. All methods were trained on the same supervised data (our train set) and tested on the evaluation set.

Metric	EdgeNet (iter 0)	NGC (iter 2)	EdgeNet (iter 2)	CCT (supervised)	CCT (semi-supervised)
mIOU	0.484	**0.526**	0.516	0.353	0.353
Accuracy	0.9001	0.9245	**0.9283**	0.8463	0.8503

Table 11: Semantic segmentation comparisons on our evaluation set with the semi-supervised CCT[47].

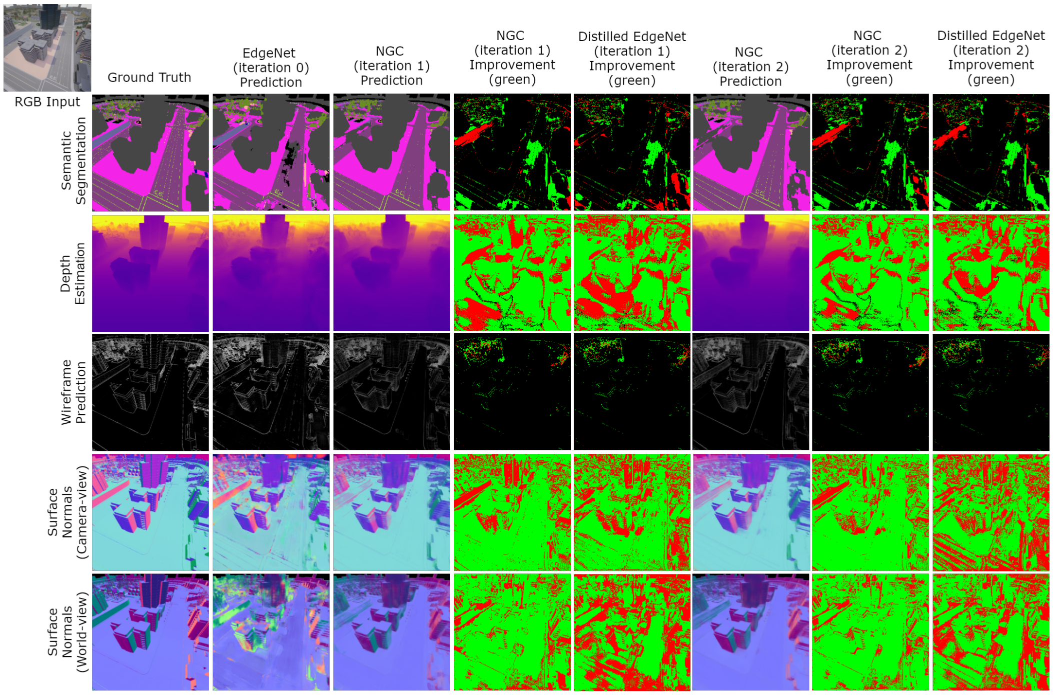

RGB Input
Ground Truth
EdgeNet (iteration 0) Prediction
NGC (iteration 1) Prediction
NGC (iteration 1) Improvement (green)
Distilled EdgeNet (iteration 1) Improvement (green)
NGC (iteration 2) Prediction
NGC (iteration 2) Improvement (green)
Distilled EdgeNet (iteration 2) Improvement (green)
Semantic Segmentation
Depth Estimation
Wireframe Prediction
Surface Normals (Camera-view)
Surface Normals (World-view)

Figure 7: Visual comparison of NGC results with baseline methods for various scene understanding tasks

4.4 Conclusions

We presented SafeUAV-Net for depth and safe landing area estimation, and the Neural Graph Consensus model for multi-task semi-supervised learning. Both approaches show promising results on synthetic and real data. Future work could explore using spatial and temporal continuity in video sequences for more robust predictions, and improving safety through visual geolocalization.

5 Towards efficient building of the 3D structure

We propose an efficient method for unsupervised learning of metric depth estimation from a single image in the context of unconstrained videos captured from UAVs. We combine the accuracy of an analytical solution based on odometry with the power of deep learning. Our approach, called UFODepth, outperforms state-of-the-art methods on a UAV dataset that we significantly extend.

5.1 UFODepth: Unsupervised learning with flow-based odometry optimization for metric depth estimation

5.1.1 Introduction

Unsupervised learning of metric depth estimation from video and odometry can have a strong impact on autonomous navigation. For UAVs, safe landing is a crucial task that should be done with high precision using solely sensors available on-board. Our approach combines two complementary directions:

UFODepth has an analytical solution for depth estimation from optical flow and measured camera velocities - corrected with a novel optimization approach. The analytical depth is then used both as input and as an additional cost term for training the final UFODepth net for depth estimation.

5.1.2 Scientific context

Recent real-time SLAM methods result in low resolution output (256x192 pixels) [48]. Other approaches that achieve real-time 3D reconstruction (e.g. NeuralRecon [49]) are based on SDF representations that do not directly output a depth map. Recent deep learning methods such as consistent depth estimation [50] or NeRF-based approaches [51, 52] produce metric depth only if they start from scaled intermediary SfM results and need fine-tuning on each scene. Several methods rely on unsupervised learning [53–55]. The two most similar works for unsupervised depth claiming real-time operation are [56] and [57]. Other recent approaches providing high-quality depth output generally require heavy pre-training on very large datasets [58, 59].

5.1.3 Flow-based odometry optimization

We aim to compute robust metric depth from optical flow and camera odometry. First we derive an analytical solution, which we then use to correct the initial odometric measurements. The image-to-image optical flow caused by camera movement can be written as a sum of two components: the linear flow $\mathbf{F}_\nu$ caused by translations, and the rotational flow $\mathbf{F}_\omega$ produced by rotation. The temporal derivative of a 3D scene point $P = (X, Y, Z)$ in the camera coordinate system is related to the camera movement by linear (ν) and angular (ω) velocities:

$$\dot{\mathbf{P}} = -\omega \times \mathbf{P} - \nu. \tag{1}$$

We can define optical flow as a function of instantaneous camera motion and depth:

$$\left(\dot{\bar{u}}\nu \;\; \dot{\bar{v}}\nu\right) = \frac{1}{Z}\left(-f \quad 0 \quad \bar{u}\; 0 \quad -f \quad \bar{v}\right)\left(\nu_x \; \nu_y \; \nu_z\ \right). \tag{2}$$

$$\left(\dot{\bar{u}}\omega \;\; \dot{\bar{v}}\omega\right) = \left(\frac{\bar{u}\bar{v}}{f} \quad -\frac{f^2+\bar{u}^2}{f} \quad \bar{v} \; \frac{f^2+\bar{v}^2}{f} \quad -\frac{\bar{u}\bar{v}}{f} \quad -\bar{u}\right)\left(\omega_x \; \omega_y \; \omega_z\right). \tag{3}$$

5.1.4 Unsupervised training

Our training procedure is inspired by [55]. We build upon the geometry consistency loss, termed $\mathcal{L}GC$, and introduce two additional losses, $\mathcal{L}Depth$ and $\mathcal{L}_{Pose}$. Our objective is minimizing:

$$\mathcal{L} = \mathcal{L}Reconstruction + \mathcal{L}GC + \mathcal{L}Depth + \mathcal{L}Pose. \tag{4}$$

An overview of our approach is shown in Figure 8.

5.1.5 Experimental analysis

We test on and extend a recently published dataset for UAV vision research. **Slanic and Herculane Dataset:** This dataset includes a total of 20 minutes of 4K video sequences from two urban scenes from Eastern Europe [56]. It includes odometry information at 10 Hz frequency. **Extended Odometry Dataset:** We introduce three novel scenes (Oveselu, Olanesti, Chilia) with various landscapes. The dataset has a total of 33 minutes of real drone flight, at 30 FPS.

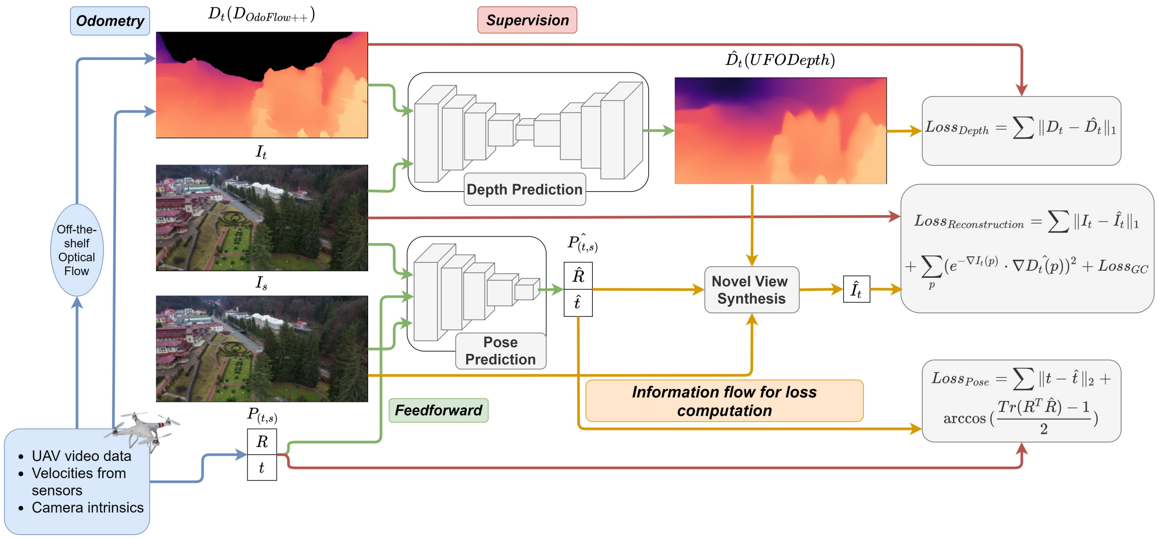

Odometry
$D_t(D_{OdoFlow++})$
Supervision
$\hat{D}_t(UFODepth)$
Depth Prediction
$Loss_{Depth} = \sum \|D_t - \hat{D}_t\|_1$
I_t
I_s
Off-the-shelf Optical Flow
Pose Prediction
$\hat{P_{(t,s)}}$
$\hat{R}$
$\hat{t}$
Novel View Synthesis
$\hat{I}_t$
$Loss_{Reconstruction} = \sum \|I_t - \hat{I}_t\|_1$
$+ \sum_p (e^{-\nabla I_t(p)} \cdot \nabla \hat{D_t}(p))^2 + Loss_{GC}$
$Loss_{Pose} = \sum \|t - \hat{t}\|_2 +$
$\arccos(\frac{Tr(R^T \hat{R}) - 1}{2})$
Information flow for loss computation
Feedforward
$P_{(t,s)}$
R
t
UAV video data
Velocities from sensors
Camera intrinsics

Figure 8: Overview of our approach (UFODepth). We combine three types of losses with an improved mathematical formulation for depth from optical flow.

Table 12: Mean absolute and relative errors against D_{SfM} ground truth depth.

Method	Slanic		Chilia		Olanesti		Herculane		Oveselu		Overall		
	Metric	Rel.	Metric	Rel.	Metric	Rel.	Metric	Rel.	Metric	Rel.	Metric	Rel.	Coverage
$D_{Triangulation}$[60]	25.39	18.9	14.79	11.2	51.82	49.3	23.19	13.4	21.79	19.9	21.29	15.6	74.0 %
$D_{OdoFlow}$ [56]	18.33	13.7	**12.90**	11.0	13.5	**11.4**	17.36	9.6	19.50	17.7	16.32	12.7	74.0 %
$D_{OdoFlow++}$	**16.06**	**12.4**	13.92	**10.4**	**13.4**	11.7	**14.46**	**8.6**	**18.52**	**17.3**	**15.27**	**12.0**	74.0 %

5.1.6 Results

Table 12 shows the performance of our analytical depth estimation method $D_{OdoFlow++}$ compared to previous work. Table 13 compares our full UFODepth approach to state-of-the-art methods. Our results show that UFODepth outperforms previous methods on average across all scenes. The analytical $D_{OdoFlow++}$ component provides accurate depth estimates where valid. The full UFODepth approach generalizes well to novel scenes while maintaining competitive inference speed.

5.2 Depth distillation: unsupervised metric depth estimation for UAVs

We also explore a depth distillation approach that combines analytical and data-driven methods. An overview is shown in Figure 9. We use three types of depth maps:

D_{SfM}: Depth from structure-from-motion, used as ground truth for evaluation. $D_{OdoFlow}$: Analytical depth from optical flow and odometry. D_{Unsup}: Depth from an unsupervised deep network, scaled to be metric using $D_{OdoFlow}$.

Table 13: Mean absolute and relative errors against D_{SfM} ground truth depth. Methods that do not provide metric estimations are scaled towards $D_{OdoFlow++}$ for fair comparison.

Method	Slanic		Chilia		Olanesti		Herculane		Oveselu		Overall	
	Metric	Relative	Metric	Relative	Metric	Relative	Metric	Relative	Metric	Relative	Metric	Relative
D_{Unsup}[55]	25.00	15.4	44.52	23.4	25.85	18.4	34.40	16.0	31.10	22.3	32.17	19.1
$D_{Ensemble}$[56]	24.83	14.9	37.93	18.4	22.46	**15.2**	34.28	16.6	33.15	22.1	30.53	17.44
Tiny-16 [56]	26.34	16.7	46.11	23.7	26.76	19.5	41.30	19.8	32.76	23.8	34.65	20.7
DPT [59]	34.33	22.8	**23.87**	**13.9**	26.36	20.1	30.48	14.8	28.57	26.8	28.72	19.7
BMD [58]	42.09	33.1	46.83	36.5	33.75	31.4	80.44	44.1	38.83	41.4	48.4	37.3
PackNet [57]	34.36	21.4	43.82	25.4	31.34	22.9	42.64	20.1	33.41	25.2	37.11	23.0
UFODepth-RGB	**21.52**	**14.4**	49.90	27.6	25.28	18.5	32.52	16.2	30.80	23.0	32.0	19.9
UFODepth	22.36	14.9	33.56	17.0	**21.98**	15.4	**26.45**	**13.0**	**26.73**	**19.4**	**26.21**	**15.9**

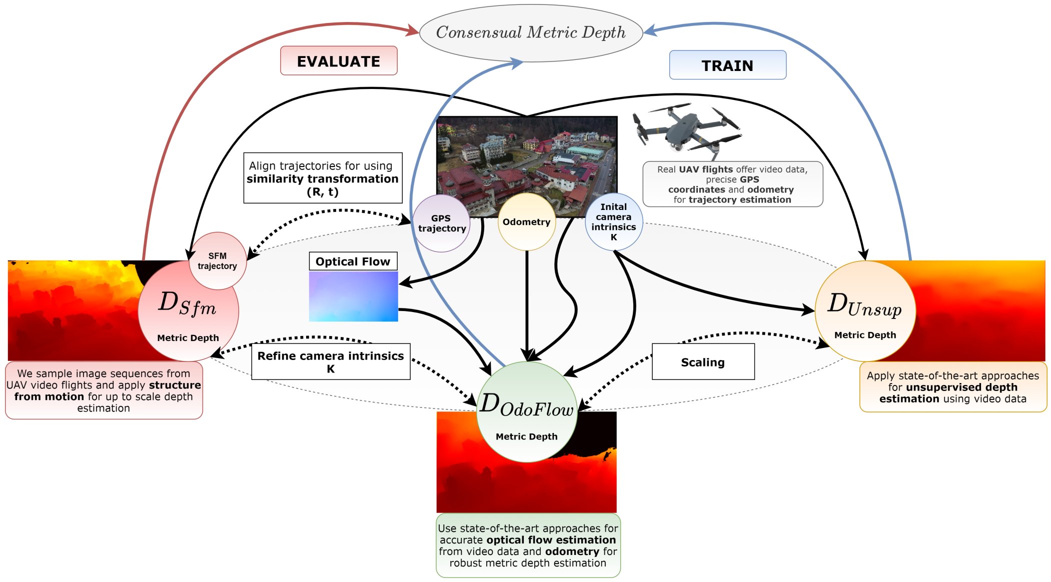

Consensual Metric Depth
EVALUATE
TRAIN
Align trajectories for using **similarity transformation** (R, t)
Real **UAV flights** offer video data, precise **GPS** **coordinates** and **odometry** for **trajectory estimation**
GPS trajectory
Odometry
Inital camera intrinsics K
SFM trajectory
Optical Flow
D_{Sfm}
Metric Depth
D_{Unsup}
Metric Depth
Refine camera intrinsics K
Scaling
$D_{OdoFlow}$
Metric Depth
We sample image sequences from UAV video flights and apply **structure from motion** for up to scale depth estimation
Apply state-of-the-art approaches for **unsupervised depth estimation** using video data
Use state-of-the-art approaches for accurate **optical flow estimation** from video data and **odometry** for robust metric depth estimation

Figure 9: Overview of our depth distillation approach, combining several complementary pathways for accurate metric depth estimation.

	Slanic		Herculane	
	Metric	Relative	Metric	Relative
D_{Unsup}	27.28 m	17.10 %	44.39 m	20.29 %
$D_{OdoFlow}$	26.05 m	16.34 %	**39.67 m**	**17.53 %**
$D_{Ensemble}$	25.63 m	15.88 %	41.18 m	18.29 %
$Tiny - 16$	**21.58 m**	**14.58 %**	46.77 m	24.09 %
$Large - 16$	21.84 m	14.65%	48.00 m	23.97 %

Table 14: Mean absolute and relative errors on entire valid map against D_{SfM} ground truth depth.

We form an ensemble teacher by CopyRetryClaude's response was limited as it hit the maximum length allowed at this time.IPlsea continueEditcombining $D_{OdoFlow}$ and D_{Unsup}, and use this to train a lightweight student network for real-time depth estimation.

5.2.1 Results

Tables 14 and 15 show results on the full image area and the "good" area where $D_{OdoFlow}$ is valid.

The distilled student networks (Tiny-16 and Large-16) outperform their teachers on the Slanic test set. While performance degrades on the unseen Herculane dataset, the students remain competitive. Table 16 shows inference speed on desktop and embedded GPUs.

	Slanic		Herculane	
	Metric	Relative	Metric	Relative
D_{Unsup}	21.06 m	15.31 %	31.61 m	16.60 %
$D_{OdoFlow}$	19.56 m	14.39 %	**24.97 m**	**12.72 %**
$D_{Ensemble}$	19.03 m	13.81 %	27.10 m	13.79 %
$Tiny - 16$	**16.11 m**	**12.90 %**	37.42 m	22.95 %
$Large - 16$	16.66 m	13.41 %	37.43 m	22.41 %

Table 15: Absolute and relative errors on the good area, where both D_{SfM} and $D_{OdoFlow}$ predictions are valid.

Method	Parameters	Desktop[FPS]	Embedded[FPS]
D_{Unsup}[55]	14,842,236	166.703 ± 3.27	11.631 ± 0.55)
$OpticalFlow(only)$[61]	15,263,888	83.404 ± 2.65	43.699 ± 3.13
$D_{OdoFlow}$	n/a	26.807 ± 6.17	10.127 ± 0.73
$Tiny - 16$[62]	1,119,862	54.922 ± 1.33	10.357 ± 0.22
$Large - 16$[62]	2,005,239	51.969 ± 1.32	9.045 ± 0.13

Table 16: Frames per second on desktop and embedded GPUs. The depth from flow algorithm runs on CPU. The desktop features a RTX 2080 GPU and Ryzen 7 3700 CPU.

5.3 Conclusions

We propose two complementary approaches for unsupervised metric depth estimation from UAV videos:

UFODepth combines an analytical depth estimation method with unsupervised deep learning, achieving state-of-the-art results on our extended UAV dataset. A depth distillation approach that uses an ensemble of analytical and unsupervised methods as a teacher to train a lightweight student network.

Both methods demonstrate good generalization to novel scenes and near real-time performance on embedded platforms, making them suitable for on-board deployment on UAVs. Future work could explore incorporating additional geometric constraints and optimizing for real-time performance on embedded devices.

6 A self-supervised cyclic neural-analytic approach for novel view synthesis and 3D reconstruction

Generating novel views from recorded videos is crucial for enabling autonomous UAV navigation. Recent advancements in neural rendering have facilitated rapid development of methods capable of rendering new trajectories. However, these methods often fail to generalize well to regions far from the training data without an optimized flight path, leading to sub-optimal reconstructions. We propose a self-supervised cyclic neural-analytic pipeline that combines high-quality neural rendering outputs with precise geometric insights from analytical methods. Our solution enhances both RGB and mesh reconstructions for novel view synthesis, particularly in undersampled areas and regions entirely distinct from the training dataset.

6.1 Introduction

Traditional approaches to novel view synthesis predominantly focus on object-centered or synthetic datasets characterized by minimal noise in the input data. Testing typically occurs on the same images used for training or at regular intervals within a sequence [63]. Although current neural rendering techniques can produce high-quality reconstructions on the training data – with PSNR values often exceeding 30 [64], camera paths diverging significantly from the training data frequently lead to inferior reconstructions. A prevalent strategy to mitigate these limitations involves segmenting datasets into smaller, consistent regions [65, 66]. However, this approach demands substantial computational resources, as a separate model must be trained for each sequence [67, 68]. We introduce a cyclic neural-analytic approach that utilizes the strengths of structure-from-motion and neural rendering methods to synthesize high-fidelity RGB images on novel poses far from the training set. Our pipeline employs a dual-phase reconstruction strategy. Initially, an analytic 3D reconstruction aligns with neural rendering branch outputs to refine geometric consistency. Subsequently, we repurpose a lightweight transformer for image restoration to achieve high-resolution reconstruction of novel 2D views. Our main contributions are:

1. We combine analytical and neural 3D reconstruction methods through a novel cyclic self-supervised transformer-based approach and show improvements for both novel view synthesis and 3D reconstruction through iterative learning.

2. Our framework demonstrates generalization capabilities, generating quality images from test locations without additional training data.

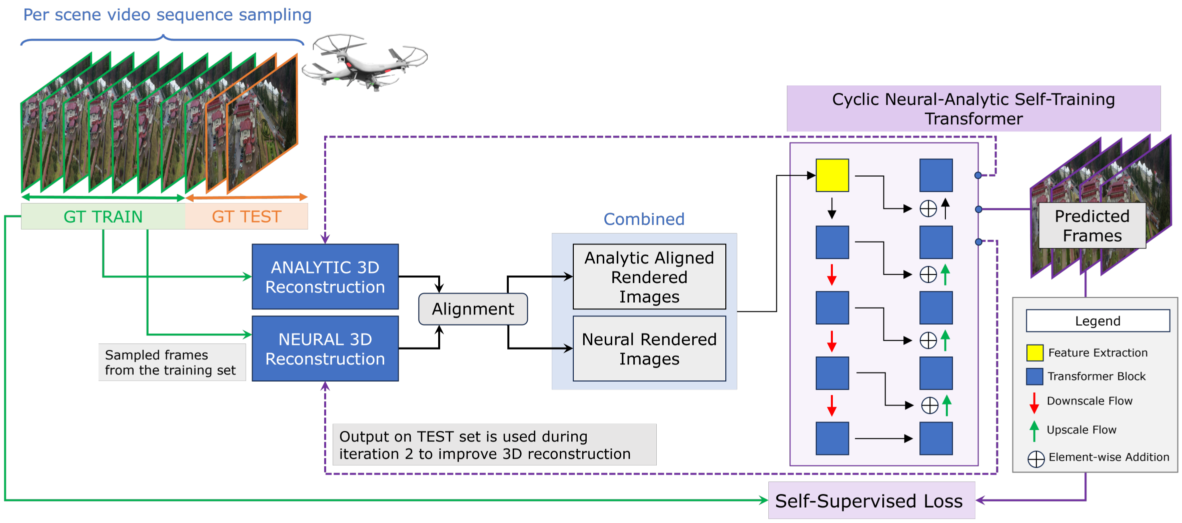

Per scene video sequence sampling
GT TRAIN
GT TEST
Cyclic Neural-Analytic Self-Training Transformer
Predicted Frames
ANALYTIC 3D Reconstruction
NEURAL 3D Reconstruction
Sampled frames from the training set
Alignment
Combined
Analytic Aligned Rendered Images
Neural Rendered Images
Output on TEST set is used during iteration 2 to improve 3D reconstruction
Legend
Feature Extraction
Transformer Block
Downscale Flow
Upscale Flow
Element-wise Addition
Self-Supervised Loss

Figure 10: Overview of our self-supervised cyclic neural-analytic pipeline for novel 2D view synthesis. We rely on both traditional and modern 3D reconstruction methods which we combine through a self-supervised transformer-based U-net style model for improved image reconstruction. We employ an iterative learning procedure in which the outputs from the first learning iteration become inputs for the next to further refine the results in terms of RGB and mesh, without additional new images. We work in the UAV video domain and use the last 20% from the image sequence as testing to simulate a more realistic reconstruction scenario. **Original RGB frames from the TEST set are used exclusively for evaluation purposes.**

3. We show improvement compared to state-of-the-art methods on difficult cases of real-world scenes captured by UAVs, spanning large areas with significant noise in pose and depth.

6.2 Method

Our pipeline consists of complementary analytical and neural modules (Figure 10). Both are capable of rendering images from novel poses. The initial iteration uses the training set for both modules. New images are generated from the test poses and included in the training data. We repeat the cycle with the newly generated images.

6.2.1 Analytic 3D Reconstruction

We employ a standard photogrammetric 3D reconstruction pipeline [14, 69, 70]. The pipeline processes unordered images to reconstruct 3D scenes, through feature extraction, matching, structure-from-motion, and multi-view stereo. The resulting dense point cloud is transformed into a textured mesh.

6.2.2 Neural 3D Reconstruction

We leverage the Gaussian Splatting method [64], which represents the scene using 3D Gaussians characterized by position, anisotropic covariance matrix, and opacity. This representation efficiently captures scene details without dense sampling. The rendering process utilizes tile-based rasterization, projecting 3D Gaussians onto a 2D plane for efficient blending.

6.2.3 Self-supervised Neural-Analytic Novel View Synthesis

We combine the previous analytical and neural 3D reconstruction by adapting a lightweight transformer-based architecture [71] for enhanced image reconstruction. The model resembles a U-net style architecture with transformer blocks used instead of simple convolutional layers. We train the model in

Table 17: Comparison to state-of-the-art methods for novel view synthesis. PSNR reconstruction results on the Aerial dataset [72] for the test set exclusively. Best numbers for each individual set are **bolded**.

MethodScene	Slanic	Olanesti	Chilia	Herculane	Mean
SFM [70]	18.43	17.63	18.75	18.27	18.27
DVGO [77]	15.34	14.34	16.43	15.43	15.38
Plenoxels [78]	17.22	15.92	18.59	16.07	15.38
Instant-NGP [13]	16.54	19.25	21.01	19.12	18.98
Neuralangelo [76]	16.79	16.20	16.96	15.82	17.29
GeoView [79]	20.28	18.64	19.04	18.17	19.03
Gaussian Splatting [80]	19.37	19.34	21.85	20.85	20.35
CNA (Ours)	**19.85**	**19.81**	**22.21**	**21.39**	**20.82**

a self-supervised manner using RGB frames from video sequences as supervision when computing pixel-wise mean squared error loss on predictions.

6.2.4 3D Cyclic Refinement

Our method includes cyclic refinement that improves results without additional supervision. By feeding rendered images from test poses back into the algorithm, we observe performance gains in image reconstruction.

6.3 Experimental Analysis

We apply our framework to a range of outdoor scenes, aiming to improve performance on the test set without additional labels or retraining. We use the following datasets:

- **Aerial dataset** [72]: Features telemetry data and diverse landscapes captured by UAVs.
- **BlendedMVS** [73]: An outdoor scene reconstruction dataset with rendered images.
- **Rubble** [74]: A drone-captured dataset with detailed imagery.
- **Tanks and Temples** [75]: We use the Family scene for object-centric evaluation.

We split each scene into 80% training and 20% testing data. Results are evaluated on the test set to ensure generalization.

6.4 Results and Discussion

We compare our method against state-of-the-art approaches including Instant-NGP [13], Neuralangelo [76], Direct Voxel Grid Optimization [77], Plenoxels [78], and Gaussian Splatting [64]. Table 6.4 shows results on the Aerial dataset. Our Cyclic Neural-Analytic (CNA) approach yields improved results without additional labels, with consistent improvements across scenes. Results on other datasets are shown in Table 6.4. We observe improvements across different scene types, including object-centric, drone-captured, and synthetically reconstructed scenes. Our ablation study (Table 6.4) shows consistent improvement across iterations, demonstrating the effectiveness of our cyclic approach. We also assess 3D reconstruction quality by comparing meshes generated from training data only versus those enhanced by CNA rendered images. Results show consistent improvement in 3D reconstruction accuracy (Table 6.4).

6.5 Conclusions

We presented a self-supervised cyclic neural-analytic pipeline that blends neural rendering with mesh-based analytical methods. Our solution enhances both RGB and 3D mesh reconstructions for novel view poses significantly different from the training set. Experiments demonstrated the effectiveness of

Table 18: Test set comparison with Instant-NGP, Gaussian Splatting and the baseline analytical reconstruction method on three additional scenes. We have selected scene *5b69cc0cb44b61786eb959bf* from BlendedMVS, the Rubble scene from MIL-19, and the Family scene for Tanks and Temples. The split structure is the same as with our previous experiments (80% training, 20% testing). Best numbers for each individual set are **bolded**.

MethodScene	BlendedMVS	Rubble	Tanks and temples	Mean
Instant-NGP [13]	14.54	11.54	14.61	13.56
Analytical [70]	13.95	13.65	15.73	14.44
Gauss.Splatt. [80]	21.52	15.53	19.63	18.89
CNA (Ours)	**21.67**	**16.04**	**19.81**	**19.17**

Table 19: Ablative studies for each of the components in our proposed method, CNA on the Aerial dataset [72]. PSNR reconstruction results are reported on the test set exclusively. (1) and (2) denote the first and second iteration of CNA. Best numbers for each individual set are **bolded**.

MethodScene	Slanic		Olanesti		Chilia		Herculane		Mean	
Iteration	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Analytic (only)	18.43	18.66	17.63	16.99	18.75	19.12	18.27	18.83	18.27	18.40
Analytic (aligned)	19.11	19.33	17.73	17.05	18.95	19.81	18.42	19.01	18.55	18.80
Neural (only)	19.49	19.73	19.48	**19.85**	21.28	22.19	21.27	21.38	20.38	20.79
CNA (Ours)	19.65	**19.85**	19.81	19.81	21.90	**22.21**	21.32	**21.39**	20.67	**20.82**

Table 20: 3D reconstruction errors over iterations on the Aerial dataset [72] for the test set exclusively. Best numbers for each individual set are **bolded**.

MethodScene	Slanic		Olanesti		Chilia		Herculane	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Iteration 1	0.6102	0.5263	0.5053	0.5013	0.5324	0.4932	0.5289	0.4802
Iteration 2	**0.4832**	**0.4392**	**0.3806**	**0.3646**	**0.3553**	**0.2934**	**0.4125**	**0.4225**

both neural and analytic modules and the usefulness of cycling. Our design allows for drop- CopyRetryClaude’s response was limited as it hit the maximum length allowed at this time.IPlease continueEditHere is the continuation and conclusion of the condensed chapter: in replacement of modules, covering new ground in this area and having the potential to push the boundaries further towards novel view synthesis grounded in the 3D physical world. While most methods are tested on synthetic scenes and on the same images used for training, we evaluated and showed improvements over state-of-the-art from viewpoints drastically different from those seen in training. Moreover, we showed that our method suffers from little degradation between seen (training) vs unseen (testing) data, unlike recent competition which shows significant degradation, indicating strong overfitting.

6.6 Future Work

We aim to improve the efficiency of the training pipeline to benefit from the latest innovations in 3D reconstruction - faster training, better accuracy, depth, surface normals and mesh as output representations. Ideally, we would iteratively improve over a textured mesh that has used all the information in the RGB image. Preprocessing for rolling shutter and motion blur could result in higher PSNR output, as shown in 3GS Deblur [81]. Comparing meshes over time is an important area of concern (e.g., detecting environmental changes such as vegetation growth or historical building decay) and we aim to adapt the concept presented here towards a mesh output instead of RGB. We would use the experience gained from synthetic data rendering and 3D reconstruction to assign pixel classes and build a more robust output. Figure 11 illustrates the consistent improvement in PSNR across test frames for our CNA method compared to baselines. This demonstrates the robustness of our

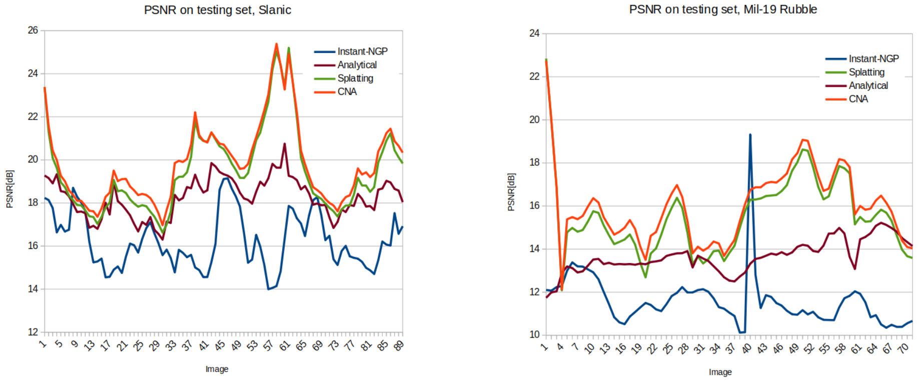


Figure 11: PSNR error on Slanic on each test frame. CNA consistently improves over the other baselines, despite not receiving additional RGB images. The performance gains are not only average, but on the vast majority of frames.

approach in generating high-quality novel views. In conclusion, our cyclic neural-analytic approach offers a promising direction for improving novel view synthesis and 3D reconstruction, particularly for challenging real-world scenes captured by UAVs. By combining the strengths of analytical and neural methods in an iterative framework, we achieve better generalization to unseen viewpoints and improved 3D reconstruction accuracy. Future work will focus on further enhancing efficiency, incorporating additional scene understanding, and extending the approach to temporal mesh comparisons for environmental monitoring applications.

7 Beyond Static Virtual Scenes: A Real-time Non-Verbal Chat for Human-AI Interaction

We present a novel approach to Human-AI interaction through real-time nonverbal chat, leveraging facial expressions and body movements to enhance engagement. Our system aims to capture and respond to user emotions using computer vision techniques, operating in real-time with minimal computational resources. We offer three complementary approaches based on retrieval, statistical, and deep learning techniques, integrating an artistic component to transmit emotions. Our experiments compare diffusion models for 2D emotion translation and introduce a 3D avatar, Maia, with facial and body movements for a more natural experience.

7.1 Introduction

Recent advances in AI, particularly Large Language Models (LLMs), have shown impressive performance in text-based tasks. However, they lack the ability to engage in complex, multi-modal interactions that characterize human communication. Current LLM-based systems are limited in their ability to tailor responses to individual needs and miss crucial nonverbal cues like gestures, body language, and facial expressions. These elements convey complex emotional and intentional information beyond written text. Our work addresses these limitations by focusing on nonverbal aspects of communication, aiming to bridge the gap between AI capabilities and the nuanced nature of human interaction. We propose a system that interprets and responds to a wide range of emotions using facial keypoints and body movements, potentially leading to more intuitive and comprehensive Human-AI communication.

7.2 Methodology

7.2.1 Nonverbal Expressions Dataset (NED)

Due to the scarcity of nonverbal expression transmission datasets, we collected our own dataset, which we will make publicly available. The dataset contains 30 videos for each emotion type, ranging from

Set	Humans (Mean)	GPT-4o
3 emotions	**91.66**	55.56
5 emotions	**50.69**	47.22

Table 21: Human and machine-level evaluation for 3 and 5 emotions generated with DreamBooth and ControlNet. Accuracy over all classes in percentages.

Method	Humans (Mean)	GPT-4o
NN	55.56	27.78
PCA	**66.67**	22.22
Retrieval	44.44	38.89
All	55.56	29.63

Table 22: Human and machine-level evaluation for each proposed 3D facial expression generation method. Accuracy over all classes in percentages.

3 to 6 seconds per video, captured at 30 FPS. For our experiments, we focused on positive emotions: happy, laughing, and surprised.

7.2.2 Introducing Maia

We created Maia, an animated character converted to 3D using VRoidStudio [82] and original artistic oil paintings. For animation, we use VSeeFace [83]. The texture used for creating our avatar is derived from a painting depicting the emotion ”happy”.

7.2.3 Evaluation Procedures

We employ two means of evaluation: an automatic evaluation using GPT-4 [84] and a human-level evaluation. For the automatic evaluation, GPT-4 analyzes individual frames from test videos, assigning emotion labels based on detected facial expressions and contextual cues. For human evaluation, we recruited individuals to evaluate the same test set as used in the automatic evaluation.

7.2.4 2D generation with diffusion

Our 2D approach uses OpenCV for video input processing, OpenPose [85] for pose estimation, and a depth estimation pipeline [86] to generate depth maps. We employ a Stable Diffusion XL model [87] with dual ControlNet [88] conditioning for image generation.

7.2.5 3D facial expression generation

We propose three methods for 3D facial response:

Expression Space Reconstruction: Uses PCA for dimensionality reduction to create an embedded space for emotions. Reaction Distillation: An unsupervised learning framework using a Teacher-Student paradigm with an MLP-based neural network. Similar Emotion Retrieval: Retrieves the most similar keypoint sequence from a predefined dataset.

7.2.6 3D body movement generation

We extended our experiments to include body movements, focusing on three emotions: ”happy to see you”, ”enthusiastic,” and ”laughing”. We collected around 100 videos of 5 seconds each for these emotions and evaluated how well Maia conveyed the desired emotion through body movement.

7.3 Results and Discussion

For the 2D generation with diffusion, human annotators achieved a mean accuracy of 91.66% in the three emotions scenario, significantly outperforming GPT-4o’s 55.56% accuracy (Table 21). For 3D facial expression generation, the PCA method performed best when evaluated by human annotators,

Set	Humans (Mean)	GPT-4o zero-shot	GPT-4o few-shot
3 emotions	**93.44**	57.67	73.00

Table 23: Human-level and automatic evaluation for the Maia body movement experiments on our test set.

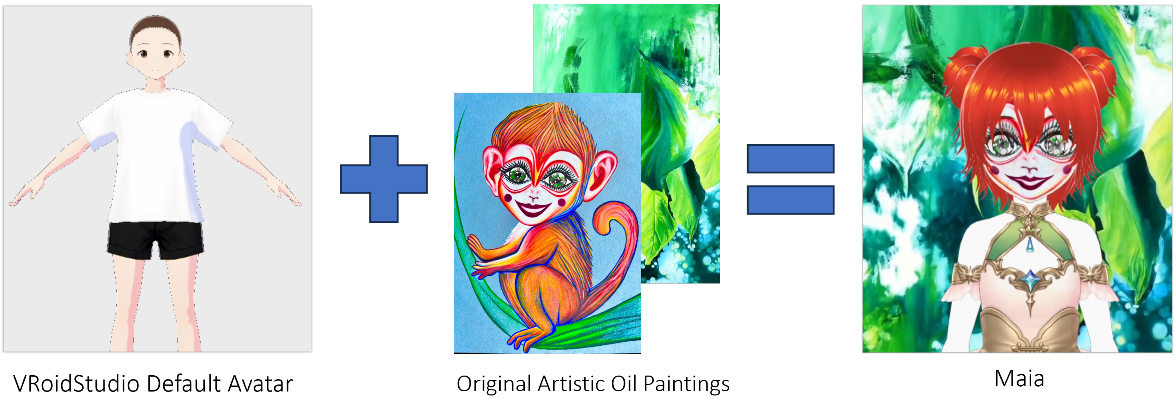


Figure 12: The process of creating our 3D avatar Maia. We used a default VRoidStudio feminine avatar on top of which we applied the facial texture from the painting on top of the avatar's face matching each correspondence manually and then customizing her with hair and clothing based on her personality. We used a different painting as the background for Maia.

while the Retrieval approach scored highest with GPT-4o (Table 22). For 3D body movement generation, human evaluators achieved a mean accuracy of 93.44%, while GPT-4o achieved 57.67% in zero-shot and 73.00% in few-shot scenarios (Table 23).

7.4 Ethical Considerations

We prioritize privacy and data protection in our approach. Our method relies on facial and body keypoints rather than raw video data, serving as a privacy-preserving feature. We extract these keypoints from facial data and emotional insights, anonymizing the information and discarding the original video input. This approach allows us to build a comprehensive data lake of emotions without risking user privacy.

7.5 Impact and Future Work

Our system has potential applications in education, therapy for individuals with communication challenges, and interactive museum exhibits. It could be particularly beneficial for special needs children who require constant engagement and interaction. Future work will focus on generating combined full body and facial keypoints-based animated avatars in real time. We also plan to explore online and continual learning procedures to enhance emotion-based human-AI interaction, involving a dynamic system that continuously updates its knowledge and behavior based on real-time interactions.

7.6 Conclusion

Our real-time nonverbal chat system represents a significant advancement in Human-AI interaction. By bridging the gap between verbal and nonverbal AI communication, we open new avenues for creating more empathetic, engaging, and naturalistic AI systems, bringing us closer to the goal of truly intuitive human-machine interfaces through art.

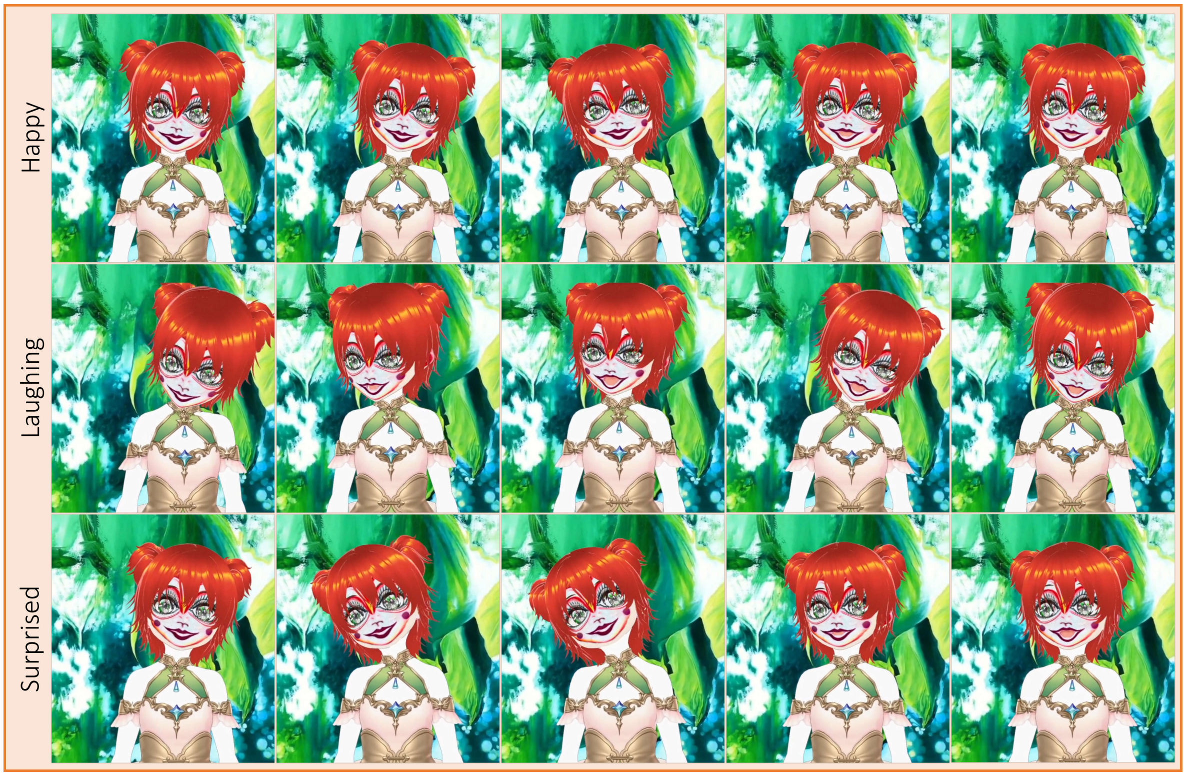


Figure 13: Maia in action. We present visual results of our PCA method applied to our 3D avatar. Currently, we focused on 3 positive emotions but can be extended to many more.

8 Conclusion

This thesis has explored several key aspects of building efficient and robust systems for aerial scene understanding, reconstruction, and novel view synthesis using unmanned aerial vehicles (UAVs). The work spans multiple interconnected areas including localization, semantic segmentation, depth estimation, 3D reconstruction, and non-verbal human-AI interaction. Throughout the various chapters, we have made several contributions that advance the state-of-the-art in these domains.

8.1 Key Insights

The insights below summarize key learnings from the thesis and also point towards directions for future research in aerial computer vision and robotics:

1. **Synergy of Geometric and Learning-Based Approaches** Combining classical geometric techniques with modern deep learning approaches often yielded robust and effective solutions. This synergy leverages the interpretability of geometric methods and the flexibility of learning-based approaches. For example, our novel view synthesis pipeline uses multi-scale voxel carving alongside a neural rendering module.

2. **Importance of Unsupervised and Self-Supervised Learning** Given the challenges of obtaining large-scale labeled data for aerial vision tasks, unsupervised and self-supervised learning techniques proved valuable. From depth estimation to novel view synthesis, we demonstrated how to leverage inherent geometric and temporal consistency in the data to train models without manual annotations.

3. **Bridging Simulation and Reality** We explored techniques for bridging the gap between synthetic training data and real-world deployment. Our work on safe landing area estimation and depth prediction showed that carefully designed synthetic datasets, combined with domain adaptation techniques, can lead to models that generalize to real-world scenarios.

4. **Efficiency and Real-Time Operation** A consistent theme was the focus on developing methods that are computationally efficient and suitable for real-time operation on embedded platforms. The resulting methods, such as SafeUAV-Net and our efficient novel view synthesis pipeline, demonstrate that it's possible to achieve competitive performance within the computational constraints of UAV platforms.

5. **Importance of Multi-Scale Representations** Across various tasks, from road detection to 3D reconstruction, we found that multi-scale representations were helpful in handling the wide range of scales present in aerial imagery. Our multi-scale voxel carving approach for novel view synthesis is an example of how this principle can be applied.

8.2 Conclusions

This thesis has explored several aspects of building systems for aerial scene understanding, reconstruction, and novel view synthesis using UAVs. The work spans multiple interconnected areas including localization, semantic segmentation, depth estimation, 3D reconstruction, and non-verbal human-AI interaction. **Chapter 2: Localization from roads and intersections in aerial images** We presented a method for automatic geolocalization of aerial images without GPS information, by matching detected roads and intersections to publicly available map data. Our approach achieved localization results even when trained on images from one city and tested on another [**?**]. Key contributions of this chapter include:

- A method for automatic geo-localization in aerial images without GPS information
- A dual-stream local-global deep CNN for detection of roads and intersections
- A geometric alignment procedure for localization and road detection enhancement

Chapter 3: Detecting roads and buildings in aerial images We introduced a dual-hop generative adversarial network (DH-GAN) for producing pixelwise segmentation of roads and intersections simultaneously at two levels of interpretation [**?**]. Key contributions of this chapter include:

- A dual-hop GAN architecture for simultaneous road and intersection detection
- A smoothing-based optimization (SBO) approach for extracting road graph structures from pixelwise segmentation
- An evaluation on a dataset of European roads, demonstrating improvements over previous methods

Chapter 4: Towards a complete understanding of the world with a drone We tackled two problems for UAV operation: safe landing area estimation and comprehensive scene understanding. For safe landing area estimation, we proposed SafeUAV-Net, a convolutional neural network for both depth and safe landing area estimation using RGB input. Key contributions include:

- A CNN architecture suitable for embedded deployment
- A synthetic dataset for training safe landing area estimation models
- Results on both synthetic data and real RGB drone footage

For comprehensive scene understanding, we introduced the Neural Graph Consensus (NGC) model, an approach to multi-task semi-supervised learning. Key aspects of this work include:

- A graph-based architecture for multi-task learning
- An approach for semi-supervised learning through consensus
- Results on multiple scene understanding tasks, including depth estimation, semantic segmentation, and pose estimation

Chapter 5: Towards efficient building of the 3D structure We introduced two contributions in the area of depth estimation from monocular imagery: First, we proposed UFODepth, a method for unsupervised learning of metric depth estimation from a single image in the context of unconstrained videos captured from UAVs. Key innovations include:

- A flow-based optimization procedure to correct noisy odometry measurements
- An analytical depth estimation method based on optical flow and corrected camera velocities
- An unsupervised learning architecture that incorporates both analytical and data-driven constraints

Second, we explored a depth distillation approach that combines multiple complementary pathways for metric depth estimation. Key aspects include:

- An ensemble approach combining analytical and learning-based depth estimation
- A distillation procedure to train a compact student network
- Generalization to novel scenes and performance on embedded hardware

Chapter 6: Putting together 3D and Novel view synthesis We addressed the problem of novel view synthesis for large-scale aerial scenes. We made two contributions in this area: First, we introduced a self-supervised approach for novel 2D view synthesis of large-scale scenes captured by UAVs. Key innovations include:

- A multi-scale geometric method based on voxel carving
- A self-supervised neural rendering module for refining the geometric reconstruction
- Performance on real-world data with noisy pose information

Second, we extended this work to a cyclic neural-analytic approach that iteratively refines both 2D novel view synthesis and 3D reconstruction quality. Key aspects of this work include:

- A cyclic refinement procedure that alternates between geometric and neural reconstruction
- Improved performance on both novel view synthesis and 3D reconstruction tasks
- Handling of large-scale scenes and noisy real-world data

Chapter 7: Beyond Static Virtual Scenes: A Real-time Non-Verbal Chat for Human-AI Interaction We explored a direction in human-AI interaction by proposing a real-time non-verbal chat system based on facial expressions. We introduced Maia, an animated avatar capable of responding to human emotions in real-time using visual cues. Key contributions of this work include:

- A dataset of nonverbal expressions for training emotion recognition and generation models
- Approaches for generating expressive responses, including 2D image generation with diffusion models and 3D facial and body animation
- An evaluation framework comparing human assessment and automatic metrics

This work opens up possibilities for AI interactions, with potential applications in education, therapy, and interactive exhibits.

8.3 Future work

While this thesis has made contributions to various aspects of aerial scene understanding and reconstruction, there are numerous directions for future research that could build upon and extend this work. We outline several promising avenues for future investigation across the different areas explored in the thesis. **Geolocalization and Mapping**

- Multi-modal fusion: Incorporate additional sensors (IMUs, low-cost GPS) for improved accuracy.

- Temporal consistency: Leverage video sequences for better stability in challenging areas.
- Adaptive map updating: Develop online algorithms for continuous map refinement.
- Semantic enrichment: Expand detection to include buildings, vegetation, and water bodies.

Semantic Segmentation and Road Extraction

- Multi-task learning: Integrate building detection and land use classification.
- Weakly-supervised learning: Utilize noisy map data for larger-scale training.
- Temporal consistency: Incorporate constraints for video sequences.
- Fine-grained road attributes: Detect lanes, road types, and traffic conditions.

Safe Landing Area Estimation

- Multi-sensor fusion: Integrate LiDAR and thermal cameras for robustness.
- Dynamic obstacle detection: Track moving vehicles and pedestrians.
- Semantic understanding: Differentiate between surface types for landing.
- Reinforcement learning: Develop end-to-end policies in simulated environments.

Depth Estimation

- Multi-view consistency: Enforce consistency across multiple frames.
- Joint optimization: Optimize depth, camera pose, and 3D structure together.
- Adaptive sampling: Focus resources on informative regions.
- Transfer learning: Develop techniques for quick adaptation to new environments.

Novel View Synthesis

- Large-scale scenes: Develop approaches for city-scale reconstruction.
- Dynamic scenes: Handle moving objects and changing lighting.
- Semantic-aware synthesis: Incorporate semantic information for improved realism.
- Real-time rendering: Push towards real-time synthesis on embedded platforms.

Human-AI Interaction

- Multimodal interaction: Incorporate gesture, pose, and physiological signals.
- Personalization: Adapt to individual users' styles over time.
- Context-aware responses: Consider conversation history and environmental factors.
- Ethical considerations: Address privacy and potential misuse concerns.

Cross-Cutting Themes

- Uncertainty estimation: Develop techniques for robust uncertainty propagation.
- Continual learning: Investigate methods for ongoing adaptation without forgetting.
- Explainability: Make complex systems more interpretable to users.
- Benchmarking: Create more comprehensive, real-world evaluation datasets.
- Energy efficiency: Develop algorithms optimized for limited power budgets.

Addressing these challenges will require interdisciplinary collaboration across computer vision, robotics, and machine learning. Advances in these areas have the potential to enable transformative UAV applications in urban monitoring, disaster response, and beyond.

References

[1] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 5007–5015. IEEE, 2015.

[2] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.

[3] Helmut Mayer, Stefan Hinz, Uwe Bacher, and Emmanuel Baltsavias. A test of automatic road extraction approaches. *International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, 36(3):209–214, 2006.

[4] Yucong Lin and Srikanth Saripalli. Road detection and tracking from aerial desert imagery. *Journal of Intelligent & Robotic Systems*, 65(1-4):345–359, 2012.

[5] Ivan Laptev, Helmut Mayer, Tony Lindeberg, Wolfgang Eckstein, Carsten Steger, and Albert Baumgartner. Automatic extraction of roads from aerial images based on scale space and snakes. *Machine Vision and Applications*, 12(1):23–31, 2000.

[6] Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *Computer Vision–ECCV 2010*, pages 210–223. Springer, 2010.

[7] Shunta Saito and Yoshimitsu Aoki. Building and road detection from large aerial imagery. In *IS&T/SPIE Electronic Imaging*, pages 94050K–94050K. International Society for Optics and Photonics, 2015.

[8] Alina Elena Marcu and Marius Leordeanu. Object contra context: Dual local-global semantic segmentation in aerial images. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[10] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[12] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[13] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022.

[14] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Kostas Karpouzis and Georgios N Yannakakis. *Emotion in games*. Springer, 2016.

[16] Angelo Cafaro, Hannes Högni Vilhjálmsson, and Timothy Bickmore. First impressions in human–agent virtual encounters. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(4): 1–40, 2016.

[17] Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620 (7976):982–987, 2023.

[18] Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. Outracing champion gran turismo drivers with deep reinforcement learning. *Nature*, 602(7896): 223–228, 2022.

[19] Dan Klang. Automatic detection of changes in road data bases using satellite imagery. *International Archives of Photogrammetry and Remote Sensing*, 32:293–298, 1998.

[20] Armin Gruen and Haihong Li. Road extraction from aerial and satellite images by dynamic programming. *ISPRS Journal of Photogrammetry and Remote Sensing*, 50(4):11–20, 1995.

[21] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[22] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007.

[23] Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on google maps street view. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 255–268. Springer, 2010.

[24] Fernando Caballero, Luis Merino, Joaquín Ferruz, and Aníbal Ollero. Unmanned aerial vehicle localization based on monocular vision and online mosaicking. *Journal of Intelligent and Robotic Systems*, 55(4-5):323–343, 2009.

[25] Xiaoming Li. A software scheme for uav's safe landing area discovery. *AASRI Procedia*, 4:230–235, 2013.

[26] Sumair Aziz, Rao Muhammad Faheem, Mudassar Bashir, Adnan Khalid, and Amanullah Yasin. Unmanned aerial vehicle emergency landing site identification system using machine vision. *Journal of Image and Graphics*, 4(1):36–41, 2016.

[27] Timo Hinzmann, Thomas Stastny, Cesar Cadena Lerma, Roland Siegwart, and Igor Gilitschenski. Free lsd: Prior-free visual landing site detection for autonomous planes. *IEEE Robotics and Automation Letters*, 2018.

[28] Maja Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1424–1445, 2000.

[29] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020.

[30] Gellert Mattyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Enhancing road maps by parsing aerial images around the world. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[31] Jonghyuk Kim and Salah Sukkarieh. Real-time implementation of airborne inertial-slam. *Robotics and Autonomous Systems*, 55(1):62–71, 2007.

[32] Fernando Caballero, Luis Merino, Joaquin Ferruz, and Aníbal Ollero. Vision-based odometry and slam for medium and high altitude flying uavs. *Journal of Intelligent and Robotic Systems*, 54 (1-3):137–161, 2009.

[33] Alina Marcu and Marius Leordeanu. Dual local-global contextual pathways for recognition in aerial imagery. *arXiv preprint arXiv:1605.05462*, 2016.

[34] Dragos Costea, Alina Marcu, Emil-Ioan Slusanschi, and Marius Leordeanu. Creating roadmaps in aerial images with generative adversarial networks and smoothing-based optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2100–2109, 2017.

[35] Per Enge Frank van Diggelen. The world's first gps mooc and worldwide laboratory using smartphones. *Proceedings of the 28th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2015)*, pages 361 – 369, 2015.

[36] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[37] Alina Marcu and Marius Leordeanu. Object contra context: Dual local-global semantic segmentation in aerial images. *AAAI Workshops*, 2017. URL https://www.aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15177/14658.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[39] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *arXiv preprint arXiv:1805.06561*, 2018.

[40] Alina Marcu, Dragos Costea, Emil Slusanschi, and Marius Leordeanu. A multi-stage multi-task neural network for aerial scene interpretation and geolocalization. *arXiv preprint arXiv:1804.01322*, 2018.

[41] Google. Google earth, 2018. URL https://www.google.com/earth/. Available at https://www.google.com/earth/, version 7.3.0.

[42] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[43] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. *arXiv preprint arXiv:1711.03938*, 2017.

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[45] Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan L Yuille. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3205–3214, 2019.

[46] Yuan Gao, Haoping Bai, Zequn Jie, Jiayi Ma, Kui Jia, and Wei Liu. Mtl-nas: Task-agnostic neural architecture search towards general-purpose multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11543–11552, 2020.

[47] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

[48] Jan Czarnowski, Tristan Laidlow, Ronald Clark, and Andrew J Davison. Deepfactors: Real-time probabilistic dense monocular slam. *IEEE Robotics and Automation Letters*, 5(2):721–728, 2020.

[49] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021.

[50] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (TOG)*, 39(4):71–1, 2020.

[51] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. *arXiv preprint arXiv:2107.02791*, 2021.

[52] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *ICCV*, 2021.

[53] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3838, 2019.

[54] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8977–8986, 2019.

[55] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in Neural Information Processing Systems*, pages 35–45, 2019.

[56] Mihai Pirvu, Victor Robu, Vlad Licaret, Dragos Costea, Alina Marcu, Emil Slusanschi, Rahul Sukthankar, and Marius Leordeanu. Depth distillation: Unsupervised metric depth estimation for uavs by finding consensus between kinematics, optical flow and deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3215–3223, June 2021.

[57] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[58] S. Mahdi, H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yağız Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proc. CVPR*, 2021.

[59] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021.

[60] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[61] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Rife: Real-time intermediate flow estimation for video frame interpolation. *arXiv preprint arXiv:2011.06294*, 2020.

[62] Alina Marcu, Dragos Costea, Vlad Licaret, Mihai Pîrvu, Emil Slusanschi, and Marius Leordeanu. Safeuav: Learning to estimate depth and safe landing areas for uavs from synthetic data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[63] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

[64] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

[65] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H. Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *CVPR*, 2023.

[66] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.

[67] Ruilong Li, Sanja Fidler, Angjoo Kanazawa, and Francis Williams. Nerf-xl: Scaling nerfs with multiple gpus, 2024.

[68] Teppei Suzuki. Fed3DGS: Scalable 3D Gaussian Splatting with Federated Learning. *arXiv preprint arXiv:2403.11460*, 2024.

[69] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[70] Carsten Griwodz, Simone Gasparini, Lilian Calvet, Pierre Gurdjos, Fabien Castan, Benoit Maujean, Gregoire De Lillo, and Yann Lanthony. Alicevision meshroom: An open-source 3d reconstruction pipeline. In *Proceedings of the 12th ACM Multimedia Systems Conference*, pages 241–247, 2021.

[71] Haiyu Zhao, Yuanbiao Gou, Boyun Li, Dezhong Peng, Jiancheng Lv, and Xi Peng. Comprehensive and delicate: An efficient transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14122–14132, 2023.

[72] Vlad Licăret, Victor Robu, Alina Marcu, Dragoş Costea, Emil Sluşanschi, Rahul Sukthankar, and Marius Leordeanu. Ufo depth: Unsupervised learning with flow-based odometry optimization for metric depth estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6526–6532. IEEE, 2022.

[73] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks, 2020.

[74] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12922–12931, June 2022.

[75] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[76] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023.

[77] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022.

[78] Fridovich-Keil and Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022.

[79] Alexandra Budisteanu, Dragos Costea, Alina Marcu, and Marius Leordeanu. Self-supervised novel 2d view synthesis of large-scale scenes with efficient multi-scale voxel carving, 2023.

[80] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

[81] Otto Seiskari, Jerry Ylilammi, Valtteri Kaatrasalo, Pekka Rantalankila, Matias Turkulainen, Juho Kannala, and Arno Solin. Gaussian splatting on the move: Blur and rolling shutter compensation for natural camera motion, 2024.

[82] Nozomi Isozaki, Shigeyoshi Ishima, Yusuke Yamada, Yutaka Obuchi, Rika Sato, and Norio Shimizu. Vroid studio: a tool for making anime-like 3d characters using your imagination. In *SIGGRAPH Asia 2021 Real-Time Live!*, pages 1–1. ACM, 2021.

[83] Vseeface. https://www.vseeface.icu/, 2023. Accessed: 2023-08-03.

[84] OpenAI. Gpt-4. https://openai.com/gpt-4, 2024. Accessed: 2024-08-25.

[85] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[86] L. Young et al. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2303.16817*, 2023.

[87] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[88] L. Zhang, Maneesh Agrawala, Frédo Durand, and X. Zhou. Controlnet: Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.