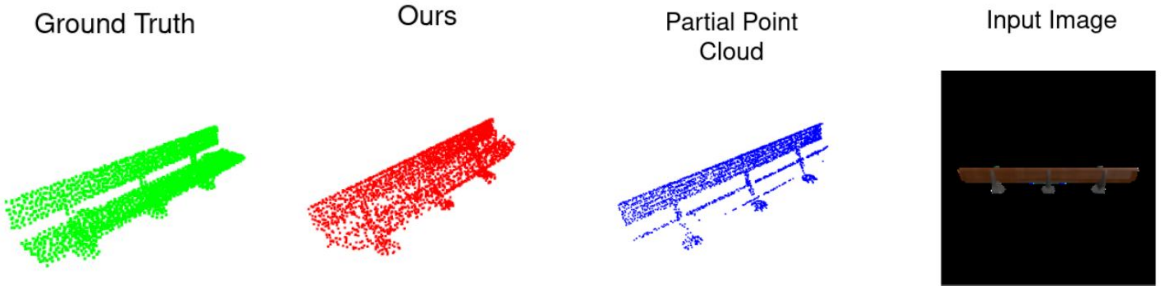
CENTER FOR DATA SCIENCE AND APPLIED MACHINE LEARNING

PES UNIVERSITY

Leveraging Pretrained Representations for Cross-Modal Point Cloud Completion

Kshitij Kale, Hrishikesh U, V sreenidhe, Shylaja S S
PES University
WACV 2025 (Oral)



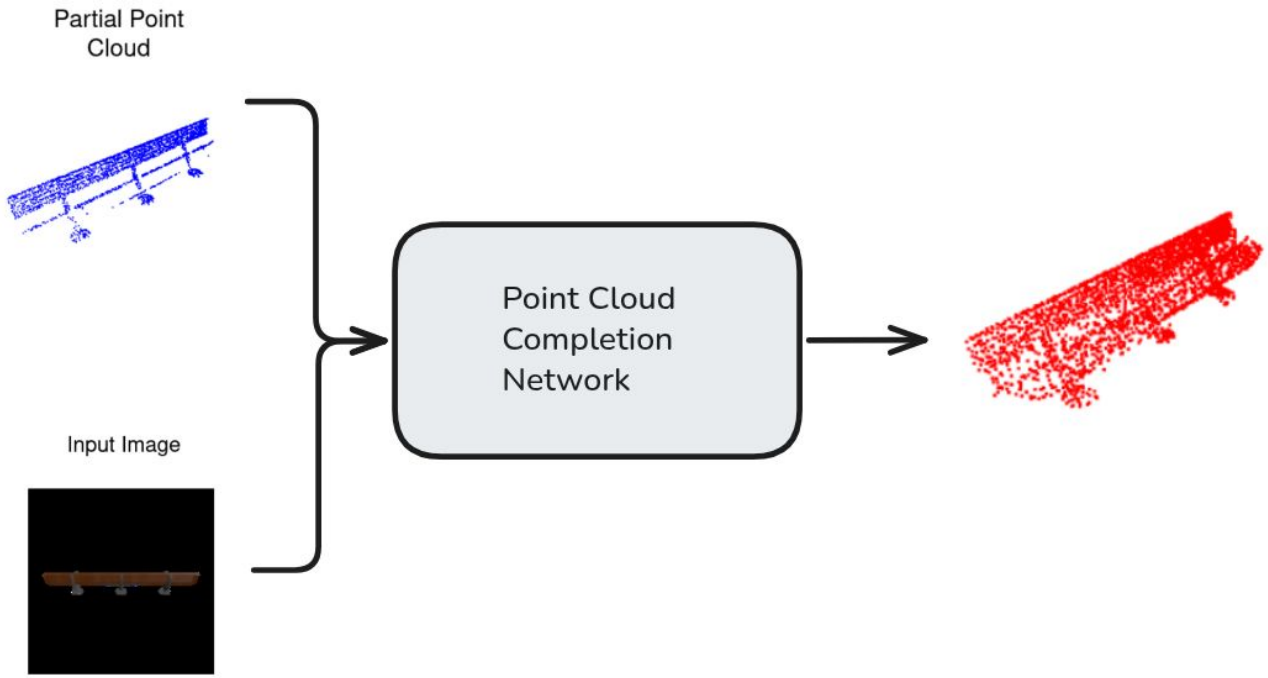
Point clouds are often incomplete

- Occlusion
- Limited viewpoints
- Sensor limitations

But many applications need full geometry for downstream tasks

Robotics • AR/VR • Autonomous driving

Goal: Point Cloud Completion

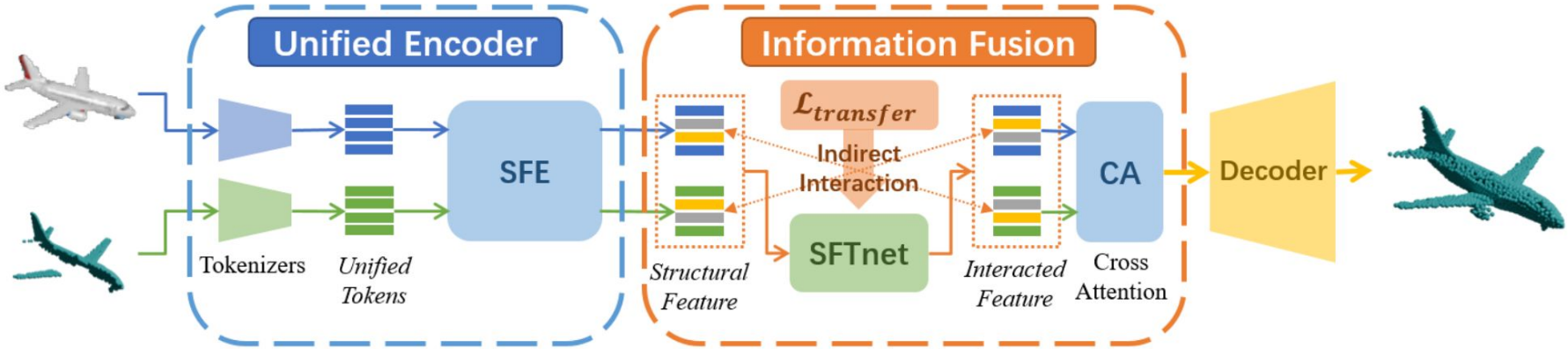


Challenge: Ill-posed problem

Related Work

Method	View-guided	Generative priors	Notes
PCN	No	No	No external priors Struggle with heavy occlusions
FoldingNet	No	No	
PoinTr	No	No	
ViPC	Yes	No	Learned from scratch Limited semantic representations
XMFNet	Yes	No	
EGIInet	Yes	No	
PCDreamer	No	Yes	Unimodal Computationally expensive
Ours	Yes	Yes	Semantic and geometric representations View-guided completion

EGIInet



- Shared Feature Extractor
- Style transfer Style Loss
- Structural loss

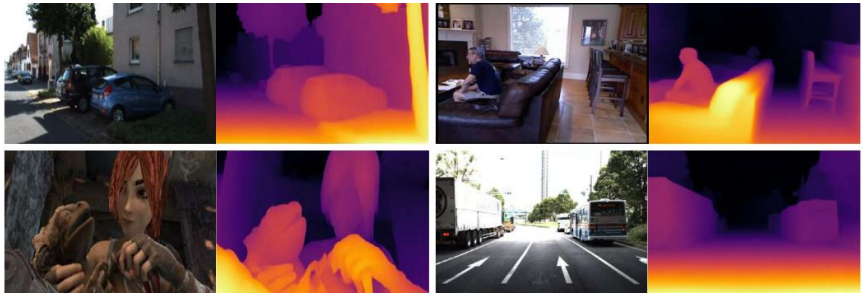
$$\mathcal{L}_{infor} = \frac{\left(\boldsymbol{G}\left(\boldsymbol{F}_{img}^{stc}\right) - \boldsymbol{G}\left(\boldsymbol{F}_{pc}^{\prime}\right)\right)^2 + \left(\boldsymbol{G}\left(\boldsymbol{F}_{pc}^{stc}\right) - \boldsymbol{G}\left(\boldsymbol{F}_{img}^{\prime}\right)\right)^2}{N \times C}$$

$$\mathcal{L}_{stc} = \left(\boldsymbol{F}_{pc}^{stc} - \boldsymbol{F}_{pc}^{\prime}\right)^2$$

Pretrained Representations

DINOv2
Self-supervised Vision Transformer
Trained on massive image datasets
Learns rich semantic representations
Encodes object category & shape priors

Depth-Anything
Large-scale depth estimator
Trained on diverse unlabeled data
Learns geometric structure cues
Produces dense depth maps



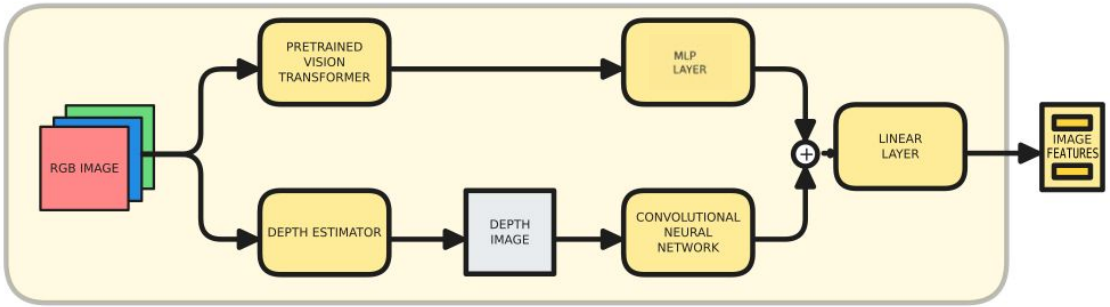
Methodology

We guide point cloud completion using pretrained semantic and geometric priors.

$$F_{geo} = \text{CNN}(\text{Depth-Anything}(I))$$

$$F_{sem} = \text{MLP}(\text{DINOv2}(I))$$

$$F_{\text{img}} = \text{Linear}(\text{Concat}(F_{\text{geo}}, F_{\text{sem}}))$$



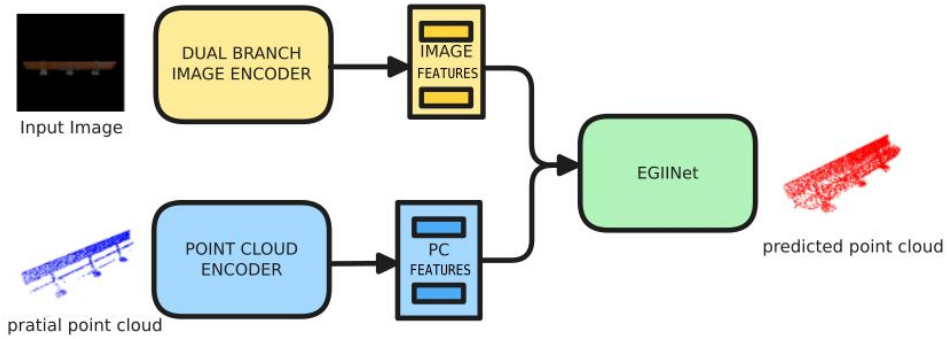
Transfer knowledge from foundation models instead of learning from scratch

Semantic → object-level understanding

Geometric → spatial structure

Priors condition the completion network to infer plausible shapes

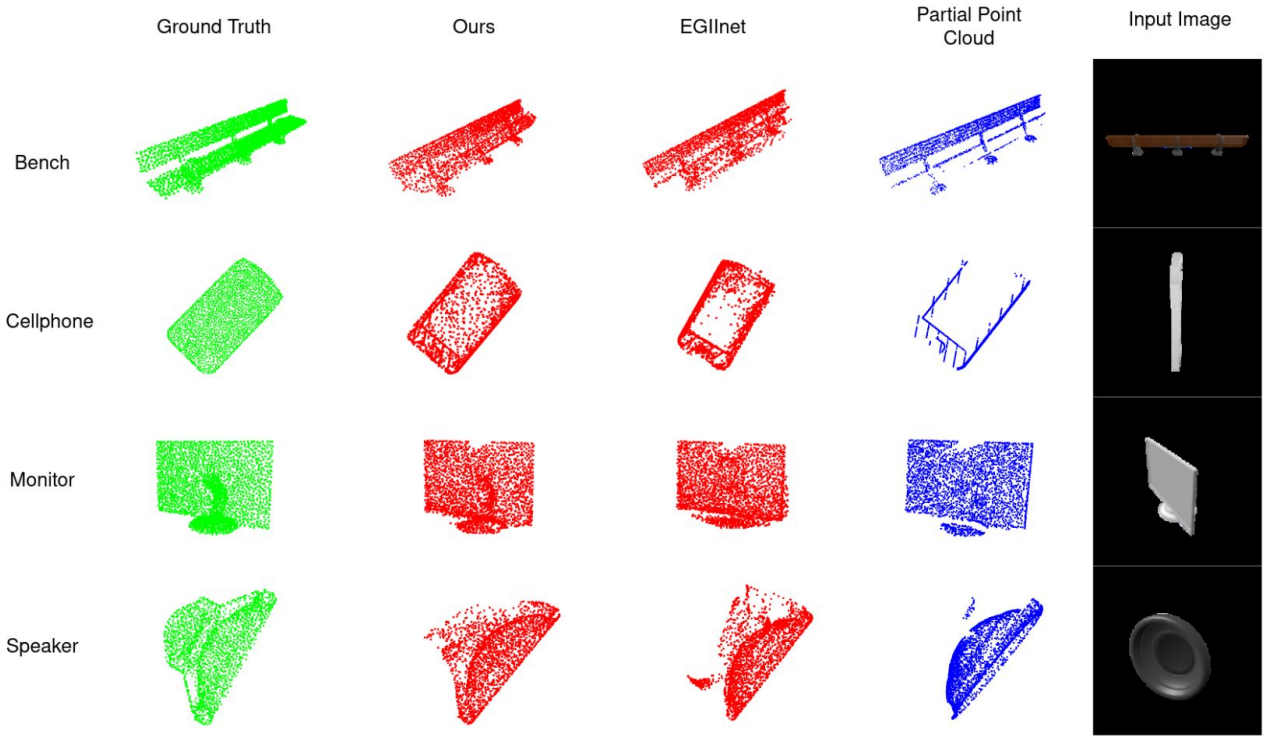
Architecture Diagram



(a) Complete Pipeline

Plug into EGIInet for +7% performance improvement

Results


Ground Truth
Ours
EGIInet
Partial Point Cloud
Input Image
Bench
Cellphone
Monitor
Speaker

Results

By integrating our Dual-branch Encoder, the Chamfer Distance is reduced across all unseen categories, achieving a 7% lower average CD compared to the baseine EGIInet

Results on unknown categories (CD $\times 10^3$ $\downarrow$(Lower is better), F-score @ 0.001$\uparrow$).

Methods	Avg		Bench		Monitor		Speaker		Cellphone	
	CD	F-score	CD	F-score	CD	F-score	CD	F-score	CD	F-score
PF-Net [8]	5.011	0.468	3.683	0.584	5.304	0.433	7.663	0.319	3.392	0.534
MSN [11]	4.684	0.533	2.613	0.706	4.818	0.527	8.259	0.291	3.047	0.607
GRNet [24]	4.096	0.548	2.367	0.711	4.102	0.537	6.493	0.376	3.422	0.569
PoinTr [28]	3.755	0.619	1.976	0.797	4.084	0.599	5.913	0.454	3.049	0.627
PointAttN [21]	3.674	0.605	2.135	0.764	3.741	0.591	5.973	0.428	2.848	0.637
SDT [31]	6.001	0.327	4.096	0.479	6.222	0.268	9.499	0.197	4.189	0.362
ViPC [32]	4.601	0.498	3.091	0.654	4.419	0.491	7.674	0.313	3.219	0.535
CSDN [12]	3.656	0.631	1.834	0.798	4.115	0.598	5.690	0.485	2.985	0.644
XMFnet [1]	2.671	0.710	1.278	0.862	2.806	0.677	4.823	0.556	1.779	0.748
EGIInet [3]	2.354	0.750	1.047	0.902	2.513	0.716	4.282	0.591	1.575	0.792
Ours	**2.192**	**0.763**	**1.011**	**0.907**	**2.255**	**0.735**	**4.039**	**0.601**	**1.466**	**0.809**

Conclusion

- Our method fuses rich Semantic Features from DINOv2 with explicit geometric data from Depth-Anything to construct a robust 3D shape prior. When integrated into the state-of-the-art EGIInet framework, our encoder yields a substantial **7% improvement in generalization performance**.

- Results highlight the effectiveness of using foundation models to overcome data limitations in specialized 3D tasks.

References

- Emanuele Aiello, Diego Valsesia, and Enrico Magli. **Cross-modal fusion for image-guided point cloud shape completion.** In *Advances in Neural Information Processing Systems*, pages 30760–30772, 2022.
- Zhaori Chen, Junsheng Li, Jiachen Zhang, and Guisong Li. **Explicitly guided information interaction network for cross-modal point cloud completion.** In *Computer Vision – ECCV 2024*, pages 535–553. Springer Nature Switzerland, 2024.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. **DINOv2: Learning robust visual features without supervision.** *arXiv preprint arXiv:2304.07193*, 2023. 1, 3, 6, 7.
- Guangshun Wei, Yuan Feng, Long Ma, Chen Wang, Yuanfeng Zhou, and Changjian Li. **PCDreamer: Point cloud completion through multi-view diffusion priors.** In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- Yifan Zhang, Yifei Chen, Si-Yuan Liu, Yong-Jin Liu, and Chang-Jian Chen. **View-guided point cloud completion.** In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2544–2553, 2021.
- Kaiyi Zhang, Ximing Yang, Yuan Wu, and Cheng Jin. **Attention-based transformation from latent features to point clouds.** In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3291–3299, 2022.
- Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. **Depth Anything: Unleashing the power of large-scale unlabeled data.** In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381, 2024. 1, 3, 6, 7.