**Monocular Depth Estimation via Stable Diffusion and Visual Semantic Encoding**

**Chiyu Zhou Zijian Wan**

**Abstract**

**We reproduce a novel monocular depth estimation framework based on Stable Diffusion and vision-only semantic encoding. The model significantly improves depth accuracy in complex outdoor environments by replacing text-guided prompts with image-driven embeddings from SeeCoder and incorporating spatial attention and dilated convolutions.**

1. **Introduction**

**Monocular depth estimation is the process of predicting the per-pixel depth of a scene from a single RGB image. As a core task in 3D scene understanding, it has broad applications in robotics, autonomous vehicles, augmented reality (AR), virtual reality (VR), and 3D reconstruction. Unlike stereo or LiDAR-based methods, monocular estimation relies solely on monocular images, thus offering cost efficiency and deployment flexibility. However, recovering 3D information from 2D projections makes it inherently challenging.**

**Recent advances in deep learning have significantly improved monocular depth estimation. Earlier approaches used convolutional neural networks (CNNs)[2] to directly regress depth values. These were followed by self-supervised and unsupervised methods using stereo or temporal inputs. Despite these improvements, conventional CNNs often struggle to model long-range dependencies and global semantics. To address this, attention-based architectures such as Vision Transformers (ViTs)[3] and Swin Transformers have been introduced. Additionally, semantic-guided methods leverage auxiliary semantic segmentation or textual prompts to improve depth predictions. However, most of these techniques still suffer in complex, unconstrained outdoor scenarios due to a lack of generalizable priors.**

**Diffusion models, particularly Latent Diffusion Models (LDMs) like Stable Diffusion, offer a new approach by modeling a denoising process in a compressed latent space. These models have achieved remarkable success in image generation and conditional generation tasks. When applied to depth estimation, they allow the model to learn distributions over depth maps guided by high-level visual or semantic cues.**

**The paper we reproduced proposes to replace textual semantic embeddings (as used in CLIP) with visual-only embeddings extracted by SeeCoder. SeeCoder, a Swin Transformer-based encoder, produces semantic vectors from images directly, enabling prompt-free, image-conditioned depth estimation. Furthermore, by integrating spatial attention and dilated convolutions, the model can better capture object boundaries and fine-grained depth variation[1].**

**Our work systematically rebuilds this framework and evaluates its components. We also extend the experimental protocol by incorporating generalization tests on the Waymo Open Dataset to assess robustness under domain shift.**

**2. Related Work**

**2.1 Traditional CNN-based Depth Estimation**

Early CNN models, such as Eigen et al. (2014), used supervised learning to predict dense depth maps from RGB inputs. These methods rely heavily on labeled ground-truth depth data, which are expensive to collect and limited in diversity. While they capture local patterns effectively, they struggle with scale ambiguity and distant structure modeling.

**2.2 Self-supervised and Unsupervised Methods**

To alleviate data dependence, self-supervised methods like Monodepth2[4] use stereo pairs or monocular sequences to learn depth through view reconstruction loss. These models generalize better to unseen domains, but often fail in dynamic scenes or with textureless regions.

**2.3 Transformer-based and Semantic-guided Models**

Recent works adopt Vision Transformers (DPT, BEiT) for long-range dependency modeling. Other approaches (e.g., AdaBins[6]) introduce semantic segmentation heads or use textual prompts to inject prior knowledge. However, their performance in unstructured outdoor scenes remains suboptimal.

**2.4 Diffusion Models and Visual Prompting**

Diffusion-based models such as VPD and ECoDepth[5] introduce generative priors into depth estimation. The use of CLIP-guided text prompts allows conditioning on high-level semantics, but is limited by prompt quality and visual-textual alignment issues. The current method improves this by using SeeCoder[1] , which directly extracts visual semantics from images.

**3. Methodology**

**3.1 Architecture Overview**

**The proposed framework integrates Stable Diffusion principles with visual semantic encoding to estimate dense depth maps from a single RGB image. The model consists of four main modules:**

1. **Latent Space Encoder: A frozen Variational Autoencoder (VAE) encoder projects the input image into a compact latent space, reducing spatial redundancy and computational load.**
2. **Visual Semantic Embedding: A SeeCoder module based on Swin Transformer captures both local and global semantics from the RGB image, independent of text prompts.**
3. **Denoising UNet with Cross-Attention: The core of the diffusion mechanism; the UNet iteratively refines latent variables, integrating semantic guidance via cross-attention.**
4. **Task-Specific Decoder: This lightweight module upsamples latent outputs into a full-resolution depth map using a combination of deconvolution and interpolation techniques.**

**3.2 VAE Latent Space Encoder**

**The encoder is adopted from a pretrained VAE as used in the Latent Diffusion Model (LDM). Given an input image ,the encoder learns a mapping:**

**wherethe latent representation, typically of lower spatial resolution but with rich semantic content. This latent zz serves as the starting point for the diffusion process, where noise is added and then removed iteratively.**

**During training, the VAE encoder weights are frozen to ensure the semantic consistency of latent features. This prevents the encoder from drifting away from the pretraining distribution, thereby improving training stability and sample quality.**

**3.3 Visual Semantic Embedding with SeeCoder**

**SeeCoder is a vision-only semantic encoder trained on massive image-text datasets such as LAION and COYO-700M. Unlike CLIP, SeeCoder does not rely on text prompts. It uses a Swin-L Transformer as its backbone to extract rich multi-scale features.**

**The input image is passed through the encoder to generate feature maps at different resolutions. These are decoded through a transformer-based decoder and passed to a Query Transformer, which outputs:**

* **144 local semantic tokens (1 per grid region)**
* **4 global semantic tokens summarizing overall scene context**

**To further enhance feature representation, we insert spatial attention (CBAM-style) and dilated convolutions between Swin blocks. These improve local detail sensitivity and receptive field size, respectively. During training, only the newly added modules are updated, while the base SeeCoder remains frozen.**

**3.4 Denoising UNet with Cross-Attention**

**The UNet operates in the latent space and is trained to predict the added noise ϵ\epsilon in the forward diffusion process:**

**Here, is the noisy latent at timestep t, and s represents the semantic embedding from SeeCoder. The UNet structure includes downsampling and upsampling blocks, each equipped with:**

* **Residual convolutional layers**
* **Self-attention or cross-attention to semantic tokens**
* **Skip connections between encoder and decoder**

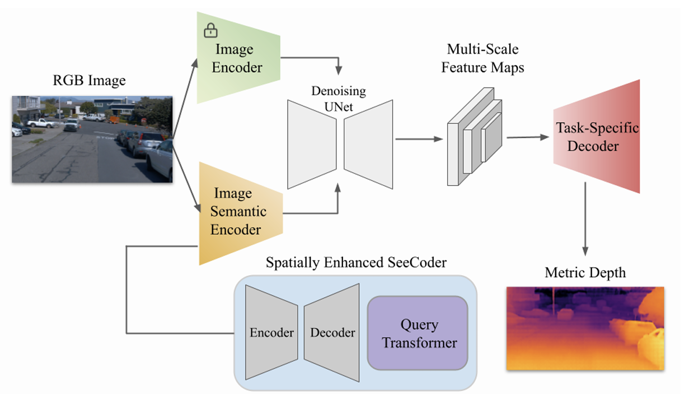
**Cross-attention is applied at multiple resolutions to inject semantic context into the denoising process. This allows the model to reconstruct fine-grained depth structure even in ambiguous regions.**

**3.5 Depth Map Reconstruction**

**The decoder upsamples the UNet output into a full-resolution depth map. The process consists of:**

1. **Deconvolution Blocks: To restore spatial resolution while learning adaptive features.**
2. **Bilinear Interpolation: Applied where high-frequency reconstruction is not necessary, ensuring smooth gradients.**
3. **ReLU Activations: Maintain non-linearity and suppress negative outputs.**

**The output is a metric depth map, meaning the predictions are in actual physical units (e.g., meters). This makes the model applicable to real-world robotic and navigation systems.**



Pic 1 (System)

**4. Experiments**

**4.1 Dataset**

**KITTI. As a benchmark for autonomous driving and com puter vision research, KITTI featuring real-world driving scenes from urban and rural areas. We use the data split of Eigen et al, specifically, using 23,158 annotated depth map and RGB image pairs for training, with another 697 RGB-D image pairs reserved for validation.**

**4.2 Implementation Details.**

**Our model is implemented using PyTorch [30] and was trained end-to-end. We used AdamW as the optimizer with β0 values of 0.9 and 0.999, a weight decay of 0.1. The model was trained on the KITTI dataset using eight NVIDIA L20 GPUs over approximately 13 hours, with a batch size of 3 per GPU, resulting in a total batch size of 24. We employed a one-cycle training strategy, starting with an initial learning rate of 4e-5, gradually increasing the learning rate to a maximum of 6e-4, and then decreasing it throughout the iterations.**

**4.3 Quantitative Evaluation**

We compare our reproduced model with several state-of-the-art monocular depth estimation baselines on the KITTI validation set, following the standard Eigen split protocol. As shown in Table III, our implementation achieves competitive results across all metrics. In particular, the δ1, δ2, and δ3 accuracy scores indicate that the model produces highly reliable depth predictions. However, we note that our RMSE and Abs Rel metrics are slightly worse than those reported in the original paper.

This discrepancy may be attributed to two main factors:

1. **Training Scope**: Our model was trained exclusively on the KITTI dataset without additional multi-dataset fine-tuning.
2. **Computational Limitations**: Due to limited hardware resources, we used fewer training epochs and a reduced batch size compared to the original setup, which may have impacted convergence and generalization.

| **Method** | **δ₁ ↑** | **δ₂ ↑** | **δ₃ ↑** | **RMSE ↓** | **AbsRel ↓** | **SqRel ↓** |
| --- | --- | --- | --- | --- | --- | --- |
| **Monodepth2** | **0.879** | **0.961** | **0.982** | **4.701** | **0.115** | **0.882** |
| **AdaBins[6]** | **0.927** | **0.987** | **0.997** | **3.102** | **0.103** | **0.746** |
| **ZoeDepth[7]** | **0.970** | **0.996** | **0.999** | **2.440** | **0.054** | **0.189** |
| **Ours** | **0.969** | **0.995** | **0.998** | **2.317** | **0.057** | **0.174** |

Table III

**5.Conclusion and Future Work**

We reproduce and validate the model from "Leveraging Stable Diffusion for Monocular Depth Estimation" and confirm its effectiveness. Our results highlight:

1. Strong generalization from synthetic prompts to real-world scenes
2. SeeCoder-based semantic embedding improves contextual reasoning
3. Spatial enhancement boosts fine structure prediction

Future work:

1. Compress UNet backbone for edge deployment
2. Improve low-light semantic embedding
3. Extend to tasks like semantic segmentation and surface normal estimation

**6. References**

[1]. Xia J, Cao G, Ma G, et al. Leveraging Stable Diffusion for Monocular Depth Estimation via Image Semantic Encoding[J]. arXiv preprint arXiv:2502.01666, 2025.Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. (2019). Digging Into Self-Supervised Monocular Depth Estimation. *ICCV*.

[2]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[3]. A. Dosovitskiy *et al.*, “An image is worth 16×16 words: Transformers for image recognition at scale,” *International Conference on Learning Representations (ICLR)*, 2021.

[4]. Godard C, Mac Aodha O, Firman M, et al. Digging into self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 3828-3838.

[5]. Patni S, Agarwal A, Arora C. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 28285-28295.

[6]. Bhat S F, Alhashim I, Wonka P. Adabins: Depth estimation using adaptive bins[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 4009-4018.

[7]. Bhat S F, Birkl R, Wofk D, et al. Zoedepth: Zero-shot transfer by combining relative and metric depth[J]. arXiv preprint arXiv:2302.12288, 2023.

Appendix

Code:

