Machine Learning Engineering & AI

Bootcamp Capstone

Step 7: Experiment With Various Models

17 April 2025

Overview:

This paper provides a comparative summary of three state-of-the-art monocular depth-estimation (MDE) models—Depth Anything V2, MiDaS, and Marigold. The comparisons focus on Performance Metrics, Loss Functions, Key Hyperparameters & Tuning, and Cross-Validation Strategy.

In practice, I did not test fine-tuning on all three models to determine the best model to use. My fine-tuning objective was to train an MDE model to recognize the following patterns:

1. If a scene has a mixture of textures (e.g. photo-realistic and cartoonish), then only the surrounding texture should have depth and the inner embedded image should be flat
2. If an embedded image overlaps a background image such that the overlapped image does not appear in the embedded image, then the embedded image must be flat (e.g. mirror or 1-D art)

In consideration of the objective patterns to fine-tune an MDE model to recognize and infer depth accordingly, my criteria for an MDE model to fine-tune was simply the model that presently seemed to be superior in inferring depth. To determine which of the three models is superior in inferring depth, I fed each model the same set of images with the following features – a mirror, indoor art, and outdoor art. I then compared the depth inference outputs from the three models. All three models had obvious flaws, with MiDaS being the inferior model among the three. A visual inspection of the depth maps favors Depth Anything v2 as a more superior model. Regardless, all three models are fooled into inferring depth for flat images within a background image, such as mirrors, indoor art, and outdoor art. Depth Anything v2 was chosen to serve as the model to be fine-tuned. The objective of this project, ultimately, is to determine an efficient manner to augment a dataset and an effective way to fine-tune an MDE model to recognize the two patterns above.

The last pages of this paper show the image set used on the three models to determine their potential infallibility in inferring depth with features of a mirror, indoor art, or outdoor art. These links are to the repository on Google Drive with the image set, and colormap depth map outputs from all three models:

1. Image set: https://drive.google.com/drive/folders/1ckWY6E46qlAar_JeCGW-F_Y4WiEVKknF?usp=drive_link
2. Depth Anything v2 outputs: https://drive.google.com/drive/folders/1BiOzwye90iFa0y9_sj31cSNEEi8quMZ2?usp=drive_link
3. MiDaS v2.1 outputs: https://drive.google.com/drive/folders/1mncaKqBON7hqzHOqPblhLnxUBSc0T-75?usp=drive_link
4. Marigold outputs: https://drive.google.com/drive/folders/19oCTw-Zc5j_wm9qTOBexNGBec9ekvGen?usp=drive_link

---

1. Performance Metrics

Monocular depth estimation is typically evaluated using metrics that capture both absolute errors and scale-invariant behavior:

- Abs Rel (mean absolute relative error)
- Sq Rel (mean squared relative error)
- RMSE and RMSE (log) (root-mean-square error in linear and log space)
- SILog (scale-invariant log RMSE)
- $\delta_1/\delta_2/\delta_3$ (percentage of predictions within thresholds 1.25, 1.25^2, 1.25^3) GitHubCSDN Blog

---

2. Loss Functions

Model	**Primary Loss(es)**
Depth Anything V2	SiLogLoss: scale-invariant log-error loss between predicted and ground-truth depth
MiDaS	L_1 or L_2 on log depth + ordinal ranking (to preserve relative ordering)
Marigold	Diffusion loss (MSE between noisy latent and predicted noise) during latent-diffusion training; optionally fine-tuned with L_1 on depth Hugging Face

---

3. Key Hyperparameters & Tuning

Model	Optimizer	LR	Batch Size	Special Tuning
Depth Anything V2	AdamW	5×10^{-6}	2–4	head LR = 10× encoder LR; weight_decay=0.01
MiDaS	Adam or AdamW	$1 \times 10^{-5} \rightarrow 1 \times 10^{-6}$	16–32	multi-scale supervision, ordinal loss weight
Marigold	DDPM (Adam)	1×10^{-4}	*diffusion step-based*	- inference steps = 100–200; guidance scale ≈ 1–8

---

4. Cross-Validation Strategy

Following is a viable cross-validation strategy to ensure robustness and guard against overfitting during fine-tuning:

1. K-fold by scene: split whole scenes (not frames) into K folds.
2. Leave-one-scene-out: train on all but one environment (e.g. outdoor vs. indoor), and test on the one environment that was excluded.
3. Continuous monitoring and exercise early stopping on validation SILog or δ_1.
4. Stratified sampling when combining multiple datasets (e.g. virtual KITTI, NYU, Hypersim).

---

5. References

- Monocular depth-estimation metrics and their definitions
  - GitHubCSDN Blog
- Depth Anything V2's use of scale-invariant log loss (SiLogLoss)
- Marigold's latent diffusion training procedure
  - https://huggingface.co/prs-eth/marigold-depth-v1-0/blob/main/README.md?utm_source=chatgpt.com
- DepthAnythingv2:
  - https://depth-anything-v2.github.io/
  - https://github.com/DepthAnything/Depth-Anything-V2/tree/main/metric_depth
- Marigold:
  - https://huggingface.co/spaces/prs-eth/marigold

    - https://github.com/prs-eth/marigold
- Midas:
    - https://pytorch.org/hub/intelisl_midas_v2/
    - https://huggingface.co/spaces/pytorch/MiDaS
    - https://github.com/isl-org/MiDaS?tab=readme-ov-file

6. Image Set to Demonstrate Infallibility

The following pages show the images used to test the fallibility of the Depth Anything v2, MiDaS v2.1, and Marigold in inferring depth with features of a mirror, indoor art, and outdoor art.


MY FANTASTIC SOCIAL MEDIA LIFE
REAL LIFE
NO WAR ONLY LOVE
CARIBBEAN VANDALS
EME.FREETHINKER


LIFE
REAL LIFE
Brillux


iStock
Credit: yacobchuk
1354175926