Depth Anything

💡 ***Monocular Depth Estimation*** ***:***
The task of estimating the depth value (distance relative to the camera) of each pixel given a single (monocular) RGB image.

💡 **Goal**

To build a foundation model for Monocular Depth Estimation capable of producing high-quality depth information for any images under any circumstances.

4. Experiment(4.4. Fine-tuned to Semantic Segmentation 까지만)

Implementation Details

- **Feature extraction**: by the DINOv2 encoder

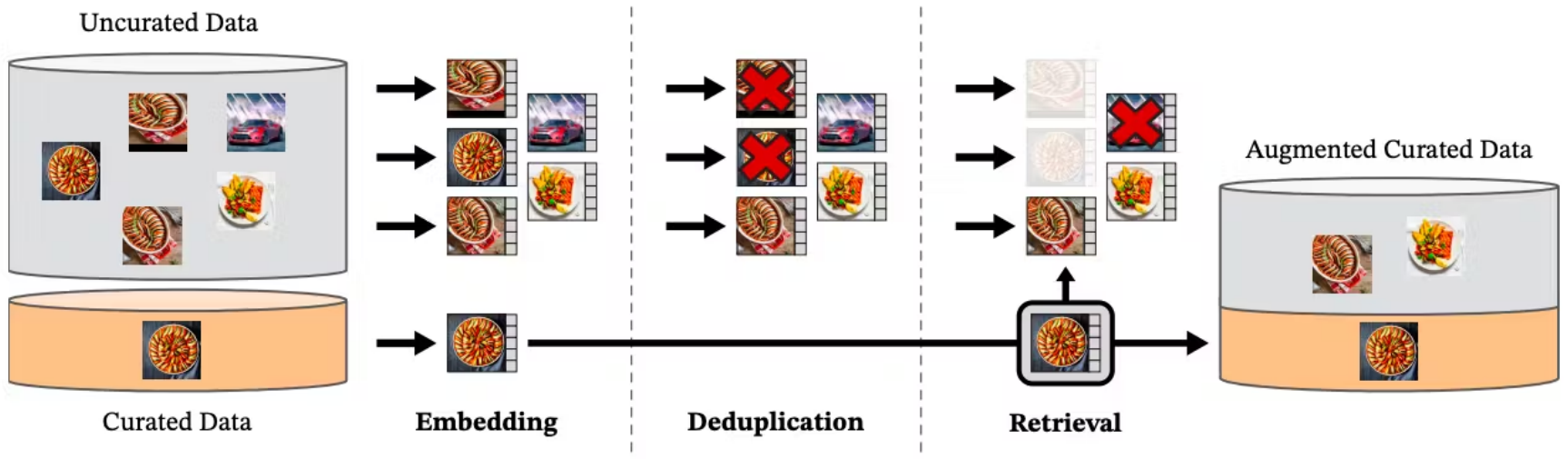
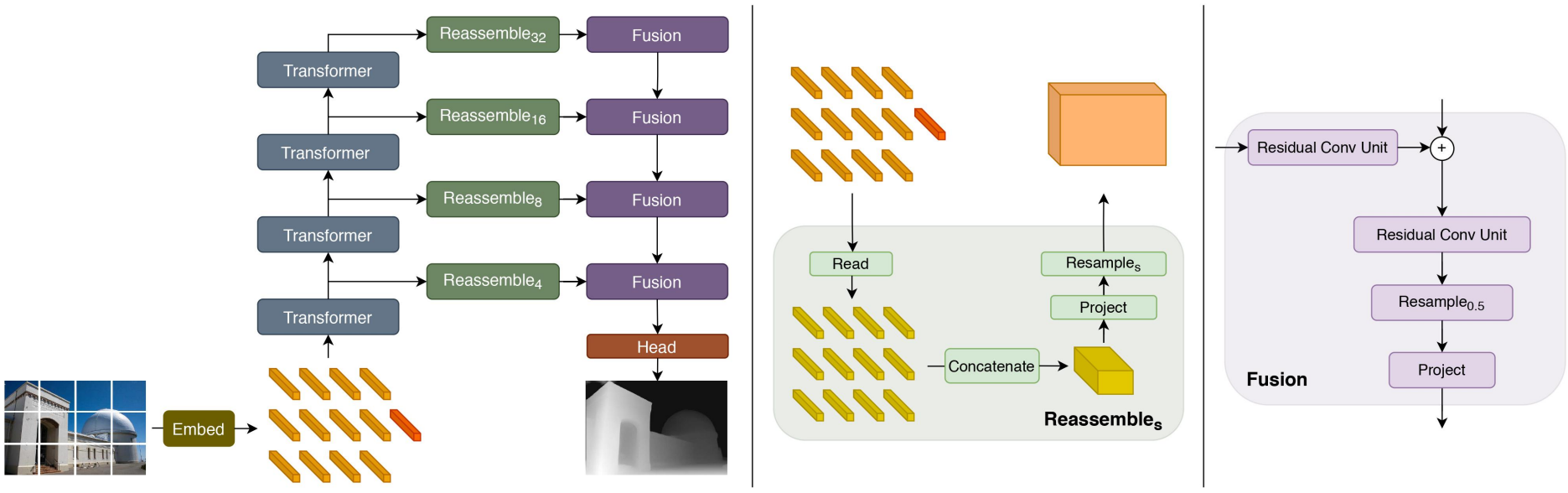


Figure 3: **Overview of our data processing pipeline.** Images from curated and uncurated data sources are first mapped to embeddings. Uncurated images are then deduplicated before being matched to curated images. The resulting combination augments the initial dataset through a self-supervised retrieval system.

- **Depth regression:** by DPT Decoer



Stage 1. Train a teacher model on **labeled images** for 20 epochs.

Stage 2. Train a student model to sweep across all **unlabeled images** for one time.

(The unlabeled images are annotated by a best-performed teacher model with a ViT-L encoder. The ratio of labeled and unlabeled images is set as 1:2 in each batch.)

Zero-shot Relative Depth Estimation

Method	Encoder	KITTI [18]		NYUv2 [55]		Sintel [7]		DDAD [20]		ETH3D [52]		DIODE [60]	
		AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
MiDaS v3.1 [5]	ViT-L	0.127	0.850	0.048	0.980	0.587	0.699	0.251	0.766	0.139	0.867	0.075	0.942
Depth Anything	ViT-S	0.080	0.936	0.053	0.972	0.464	0.739	0.247	0.768	0.127	**0.885**	0.076	0.939
	ViT-B	0.080	0.939	0.046	0.979	**0.432**	0.756	0.232	0.786	**0.126**	0.884	0.069	0.946
	ViT-L	**0.076**	**0.947**	**0.043**	**0.981**	0.458	**0.760**	**0.230**	**0.789**	0.127	0.882	**0.066**	**0.952**

Table 2. **Zero-shot relative** depth estimation. **Better:** AbsRel ↓ , δ_1 ↑. We compare with the best model from MiDaS v3.1. Note that MiDaS ***does not*** strictly follow the zero-shot evaluation on KITTI and NYUv2, because it uses their training images. We provide three model scales for different purposes, based on ViT-S (24.8M), ViT-B (97.5M), and ViT-L (335.3M), respectively. **Best**, second best results.

Better than the previously best **relative** MDE model MiDaS v3.1!

→ Insight 1) Labelled image를 더 많이 활용하는 MiDaS 모델보다도 훨씬 성능이 좋은 것을 알 수 있음

→ Insight 2) 더 큰 모델인 ViT-L을 활용한 MiDaS보다도 ViT-B 모델을 활용한 Depth Anything의 성능이 더 좋았던 경우들이 존재 (Sintel, DDAD, ETH3D)

→ Insight 3) KITTI and NYUv2의 데이터의 경우 MiDaS는 이 데이터들이 훈련 셋에 포함이 되어 zero-shot task가 아니었음에도 Depth Anything이 더 뛰어나다는 것이 파악됨

Fine-tuned to Metric Depth Estimation

> 💡 **Metric Depth Estimation이란?**
> Metric depth estimation refers to estimating depth in **real-world units** such as meters or centimeters.

실험 1) In-domain metric depth estimation: where the model is trained and **evaluated on the same** domain

Method	*Higher is better* ↑			*Lower is better* ↓		
	δ_1	δ_2	δ_3	AbsRel	RMSE	log10
AdaBins [3]	0.903	0.984	0.997	0.103	0.364	0.044
DPT [47]	0.904	0.988	0.998	0.110	0.357	0.045
P3Depth [44]	0.898	0.981	0.996	0.104	0.356	0.043
SwinV2-L [40]	0.949	0.994	0.999	0.083	0.287	0.035
AiT [42]	0.954	0.994	0.999	0.076	0.275	0.033
VPD [87]	0.964	0.995	0.999	0.069	0.254	0.030
ZoeDepth* [4]	0.951	0.994	0.999	0.077	0.282	0.033
Ours	**0.984**	**0.998**	**1.000**	**0.056**	**0.206**	**0.024**

Table 3. **Fine-tuning and evaluating on NYUv2** [55] with our pre-trained MDE encoder. We highlight **best**, second best results, as well as **most discriminative metrics**. ∗: Reproduced by us.

→ Depth Anything이 이전의 sota였던 VPD보다 큰 차이의 성능 차이로 더 좋은 성능을 보임

실험 2) Zeroshot metric depth estimation: where the model is trained on one domain but **evaluated in different domains**

(이 실험에선 pretrained encoder을 NYUv2나 KITTI에 finetuning해서 활용함)

Method	SUN RGB-D [57]		iBims-1 [29]		HyperSim [49]		Virtual KITTI 2 [8]		DIODE Outdoor [60]	
	AbsRel (↓)	δ_1 (↑)	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
ZoeDepth [4]	0.520	0.545	0.169	0.656	0.407	0.302	0.106	0.844	0.814	0.237
Depth Anything	**0.500**	**0.660**	**0.150**	**0.714**	**0.363**	**0.361**	**0.085**	**0.913**	**0.794**	**0.288**

Table 5. **Zero-shot metric** depth estimation. The first three test sets in the header are indoor scenes, while the last two are outdoor scenes Following ZoeDepth, we use the model trained on NYUv2 for indoor generalization, while use the model trained on KITTI for outdoor evaluation. For fair comparisons, we report the ZoeDepth results reproduced in our environment.

→ Test dataset이 indoor이든 outdoor이든 MiDaS encoder을 활용한 ZoeDepth보다 더 좋은 성능을 보인 것을 확인할 수 있음

실험 1 & 실험 2 결론: Depth Anything is better than the previously best **metric** MDE model ZoeDepth!

Fine-tuned to Semantic Segmentation

→ To examine the semantic capability of our MDE encoder.

→ MDE encoder을 semantic segmentation dataset에 finetuning함.

Method	Encoder	mIoU (s.s.)	m.s.
Segmenter [58]	ViT-L [16]	-	82.2
SegFormer [70]	MiT-B5 [70]	82.4	84.0
Mask2Former [12]	Swin-L [39]	83.3	84.3
OneFormer [24]	Swin-L [39]	83.0	84.4
OneFormer [24]	ConvNeXt-XL [41]	83.6	84.6
DDP [25]	ConvNeXt-L [41]	83.2	83.9
Ours	ViT-L [16]	**84.8**	**86.2**

Table 7. Transferring our MDE pre-trained encoder to **Cityscapes** for semantic segmentation. We ***do not*** use Mapillary [1] for pre-training. s.s./m.s.: single-/multi-scale evaluation.

[Table 7: Cityscapes Dataset Semantic Segmentation]

Method	Encoder	mIoU
Segmenter [58]	ViT-L [16]	51.8
SegFormer [70]	MiT-B5 [70]	51.0
Mask2Former [12]	Swin-L [39]	56.4
UperNet [69]	BEiT-L [2]	56.3
ViT-Adapter [11]	BEiT-L [2]	58.3
OneFormer [24]	Swin-L [39]	57.4
OneFormer [24]	ConNeXt-XL [41]	57.4
Ours	ViT-L [16]	**59.4**

Table 8. Transferring our MDE encoder to **ADE20K** for semantic segmentation. We use Mask2Former as our segmentation model.

[Table 8: ADE20K Dataset Semantic Segmentation]

→ Depth Anything의 Pretrained Encoder이 monocular depth estimation task와 semantic segmentation task에서 우수한 성능을 입증할 수 있었다는 점에서 middle-level이나 high-level visual perception system을 위해 **일반적인 encoder로 활용할 수 있을 것**이라고 예상함.

진짜 잘 되네요...!