DIFFUSION MODELS ARE FEW-SHOT LEARNERS FOR DENSE VISION TASKS

Anonymous authors
Paper under double-blind review

APPENDIX

A BROADER IMPACT

Our method, which employs diffusion for general few-shot dense tasks, offers significant advantages beyond technical improvements. It substantially reduces labor costs associated with pixel-by-pixel annotation of visual dense tasks, making model deployment more cost-effective and accessible, especially for resource-limited projects. Additionally, the few-shot nature of our approach reduces energy consumption, lowering the environmental impact by decreasing the need for extensive data and computational resources. This aligns with broader goals of energy conservation and emission reduction. By democratizing access to advanced machine learning technologies, our method enables smaller entities and individuals to innovate and implement AI solutions, promoting more responsible and ethical AI development.

B RESULTS ON FULLY TRAINING SET

We include the results of full training set in Table. 1. Although VPD's performance in the few-shot setting is not strong, with more training data, we can see that its performance improves significantly because it fine-tunes more parameters. In contrast, we only fine-tune the concept embeddings with a few hundred parameters. However, our method still outperforms VTM even after training on the full training set, demonstrating the higher potential of the diffusion prior. We also reported the 95% confidence interval, and it can be seen that our method, leveraging a very general prior, achieved more stable results compared to VTM.

C RESULTS WITH DIFFERENT NUMBER OF TRAINING SAMPLES

In Fig 1, we illustrate the impact of using 10, 20, 50, and 100 training samples on our method and VPD across all 12 tasks. It can be observed that our method consistently adapts better to new tasks compared to VPD when fewer than 100 training examples are provided. Moreover, as the number of training samples increases, the performance of both methods improves accordingly.

Table 1: We present the results on 10 tasks from Taskonomy and 2 tasks from NYUv2. For Taskonomy tasks, 10-shot training examples are used for each of them, and for NYU tasks, we use 20 examples. To also evaluate the statistical robustness, we run each number for 100 times and report the 95% confidence interval. Besides segmentation task, lower number indicates better performance. Our method consistently outperforms VTM on all few-shot tasks, especially on out-of-domain tasks. And our method better unleashes the power of diffusion prior for few-shot dense prediction compared to VPD.

	Few-shot			Fully Supervised		
	VTM	VPD	Ours	VTM	VPD	Ours
EucDepth	0.0812 ±0.0065	0.1056 ±0.0102	**0.0776** ±0.0072	0.0524	0.0456	0.0498
Z-depth	0.0347 ±0.0035	0.0404 ±0.0037	**0.0308** ±0.0038	0.0257	0.0210	0.0236
2DEdge	0.0818 ±0.0021	0.0965 ±0.0023	**0.0625** ±0.0022	0.0154	0.0131	0.0136
3DEdge	0.0917 ±0.0028	0.1226 ±0.0044	**0.0812** ±0.0040	0.0638	0.0564	0.0599
2DKeypoint	0.0671 ±0.0038	0.0697 ±0.0035	**0.0626** ±0.0040	0.0337	0.289	0.306
3DKeypoint	0.0512 ±0.0018	0.0670 ±0.0027	**0.0389** ±0.0014	0.0360	0.0298	0.0324
Reshading	0.1308 ±0.0058	0.1609 ±0.0044	**0.1284** ±0.0049	0.834	0.756	0.772
Curvature	0.0413 ±0.0010	0.0498 ±0.0019	**0.0376** ±0.0023	0.0345	0.0291	0.329
Normal	11.7850 ±0.4580	14.4381 ±0.3097	**10.1346** ±0.0361	6.2418	5.7963	5.9821
SemSeg	0.3980 ±0.0350	0.3484 ±0.0308	**0.4178** ±0.0361	0.4618	0.4905	0.4784
NYUDepth	0.73 ±0.09	0.49 ±0.11	**0.43** ±0.08	0.35	0.25	0.29
NYUNormal	26.1 ±3.8	18.5 ±1.7	**16.4** ±1.6	18.2	14.8	14.9

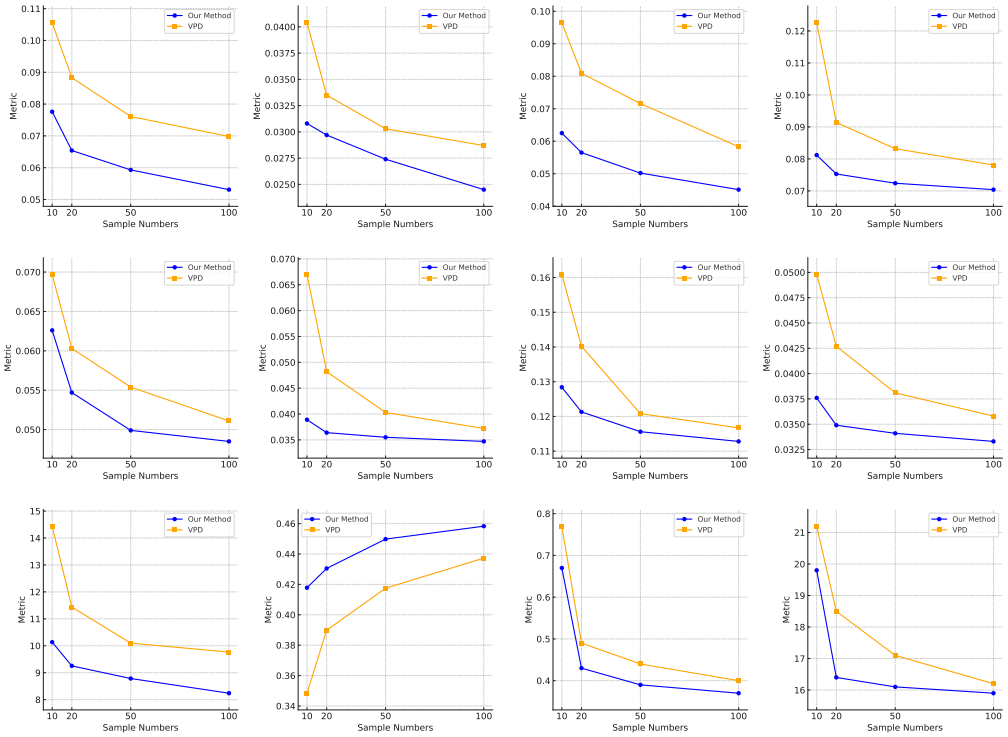

Figure 1: We present the impact of using different numbers of training samples on our method and VPD across all 12 tasks.