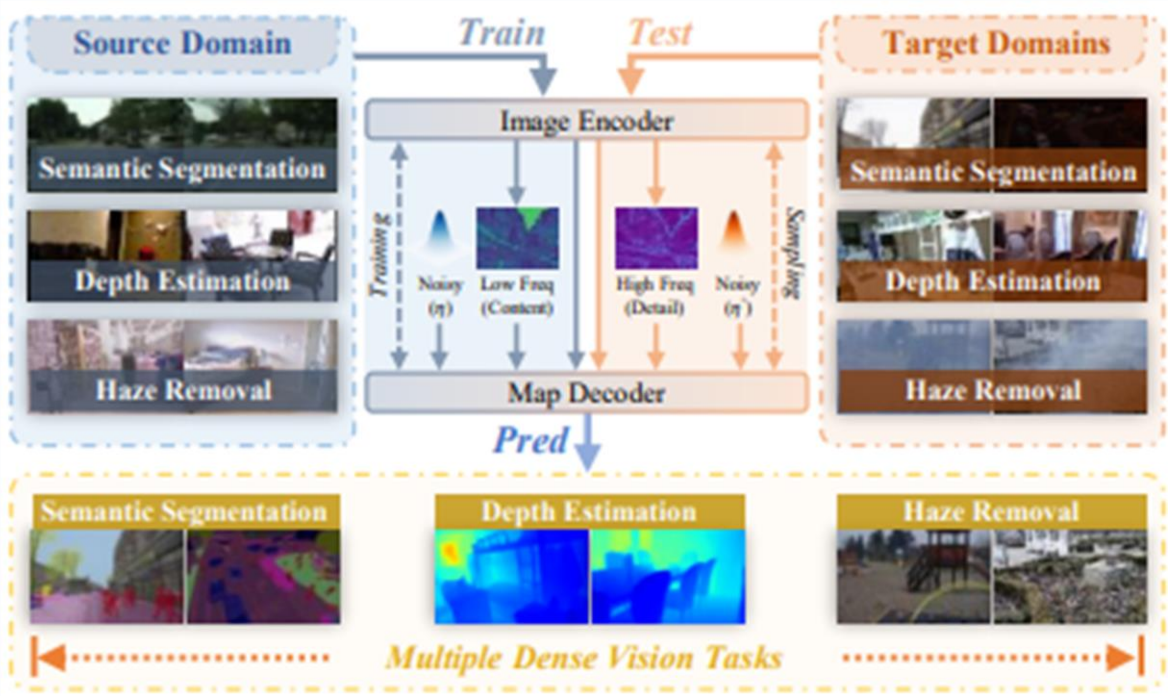
A Simple yet Mighty Hartley Diffusion Versatilist for Generalizable Dense Vision Tasks

Qi Bi, Jingjun Yi, Huimin Huang, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang, Yuexiang Li, Yefeng Zheng

[1]

Background

- Diffusion models excel across dense vision tasks but rarely generalise to unseen conditions.
- Different tasks (segmentation, depth, dehaze) exhibit disparate domain shifts..
- Motivate a solution that extracts structural content and detail cues for robust predictions.



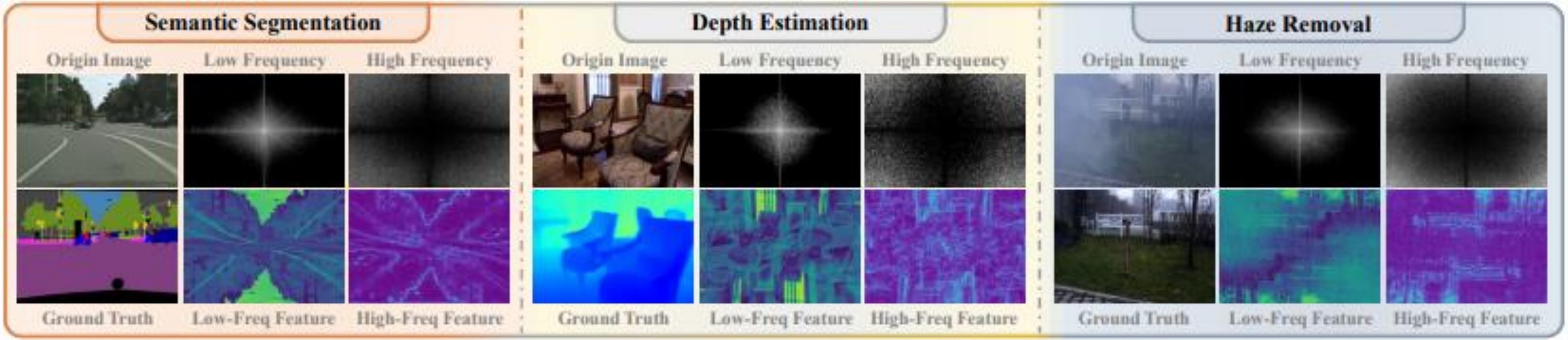
[2] [3]

Contribution

- This paper presents HarDiff, a diffusion-based method for generalizable dense vision tasks, aiming to prompt the robustness and transferability across various unseen domains.
- Low-Frequency Training Process: task-related content
- High-Frequency Sampling Process: utilizing it as a detail-oriented guidance
- Extensive experiments on three cross-domain dense pixel prediction tasks with twelve datasets show its superiority over the baselines

[2] [3]

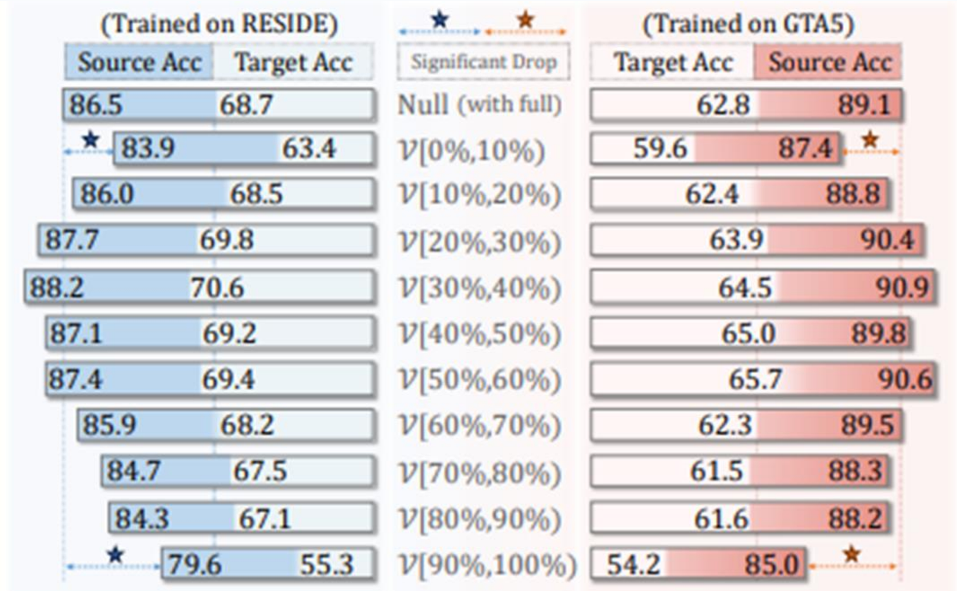
Observation



- Low-frequency bands capture global semantics (object position & shape).
- High-frequency bands capture fine details (textures & edges).
- Hartley transform operates in the real domain and fuses frequencies effortlessly.
- This dual-property motivates separate treatments for training (content) and sampling (detail).

[4] [5]

Band Response Analysis

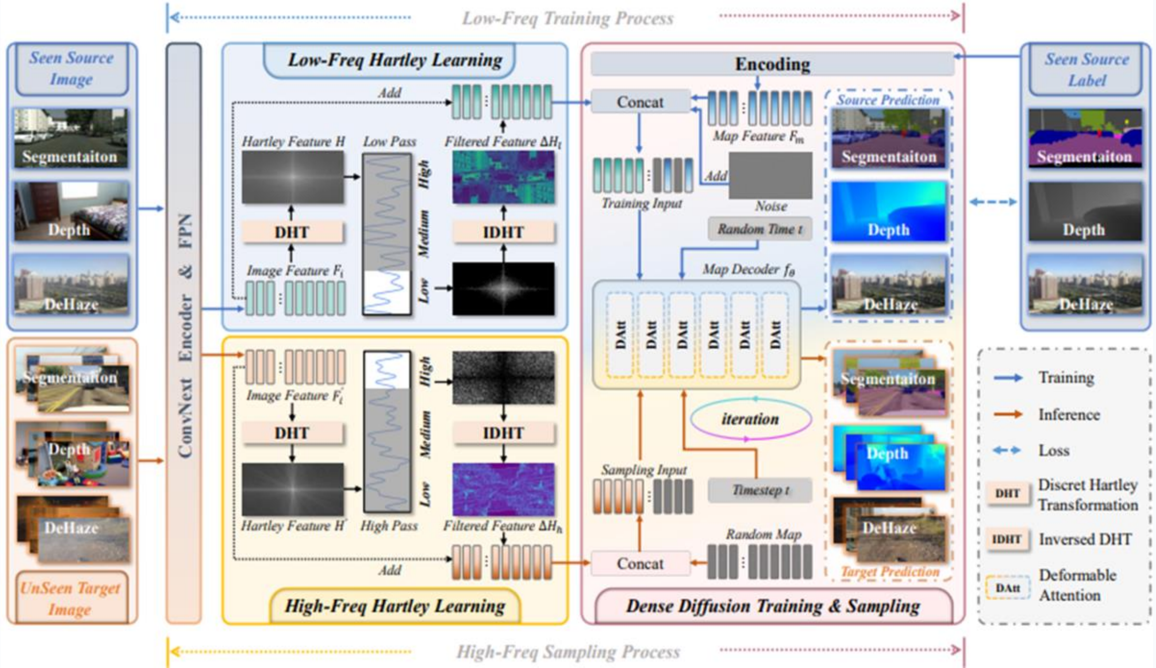


- Highest band V[0,10%) encodes detail for discriminating unseen domains.
- Lowest band V[90%,100%) encodes structural content essential for recognising in-domain samples.
- Removing either band significantly hurts accuracy on source and target domains.
- These findings inspire separate low- and high-frequency modules.

[6]

Method & Framework

- ConvNeXt encoder extracts multi-scale features.
- Low-frequency learning injects V[90%,100%] to reinforce structural priors.
- High-frequency learning injects V[0%,10%] to recover fine details.
- Unified map decoder handles segmentation, depth and dehaze within one diffusion.



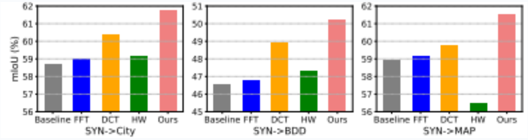
[7]

Numerical Results

- HarDiff achieves the highest mIoU across all domains.
- Average improvement of ≈2% mIoU over the second best method.
- Beyond segmentation, HarDiff attains state-of-the-art depth estimation and haze removal.

Method	Encoder	Trained on SYNTHIA (S) → C	→ B	→ M	Avg.
CNN Based:					
DRPC [91] [ICCV'2019]	ResNet-50	35.65	31.53	32.74	33.31
ISW [18] [CVPR'2021]	ResNet-50	35.83	31.62	30.84	32.76
SAW [56] [CVPR'2022]	ResNet-50	38.92	35.24	34.52	36.23
AdvStyle [98] [NeurIPS'2022]	ResNet-50	37.59	27.45	31.76	32.27
Transformer Based:					
CMFormer [9] [AAAI'2024]	Swin-B	44.59	33.44	43.25	40.43
VFM Based:					
REIN [74] [CVPR'2024]	ViT-L	48.59	44.42	48.64	47.22
SET [85] [MM'2024]	ViT-L	49.65	45.45	49.45	48.18
FADA [8] [NeurIPS'2024]	ViT-L	50.04	45.83	49.86	48.58
tqdm [54] [ECCV'2024]	ViT-L	57.99	52.43	54.87	55.10
Diffusion Based:					
PTDiffSeg [25] [ArXiv'2023]	Diffusion	49.3	-	-	-
FC-CLIP [87] [NeurIPS'2023]	ConvNeXt-L	38.0	29.9	39.0	35.6
DDP [33]* [ICCV'2023]	ConvNeXt-L	58.7	46.6	58.9	54.7
DIDEX [53] [WACV'2024]	Diffusion	59.8	47.4	59.5	55.6
CLOUDS [2] [CVPR'2024]	ConvNeXt-L	53.4	47.0	55.8	52.1
DGInStyle [34]* [ECCV'2024]	Diffusion	58.4	46.8	57.6	54.3
HarDiff (Ours)	ConvNeXt-L	**61.8**	**50.2**	**61.5**	**57.8**

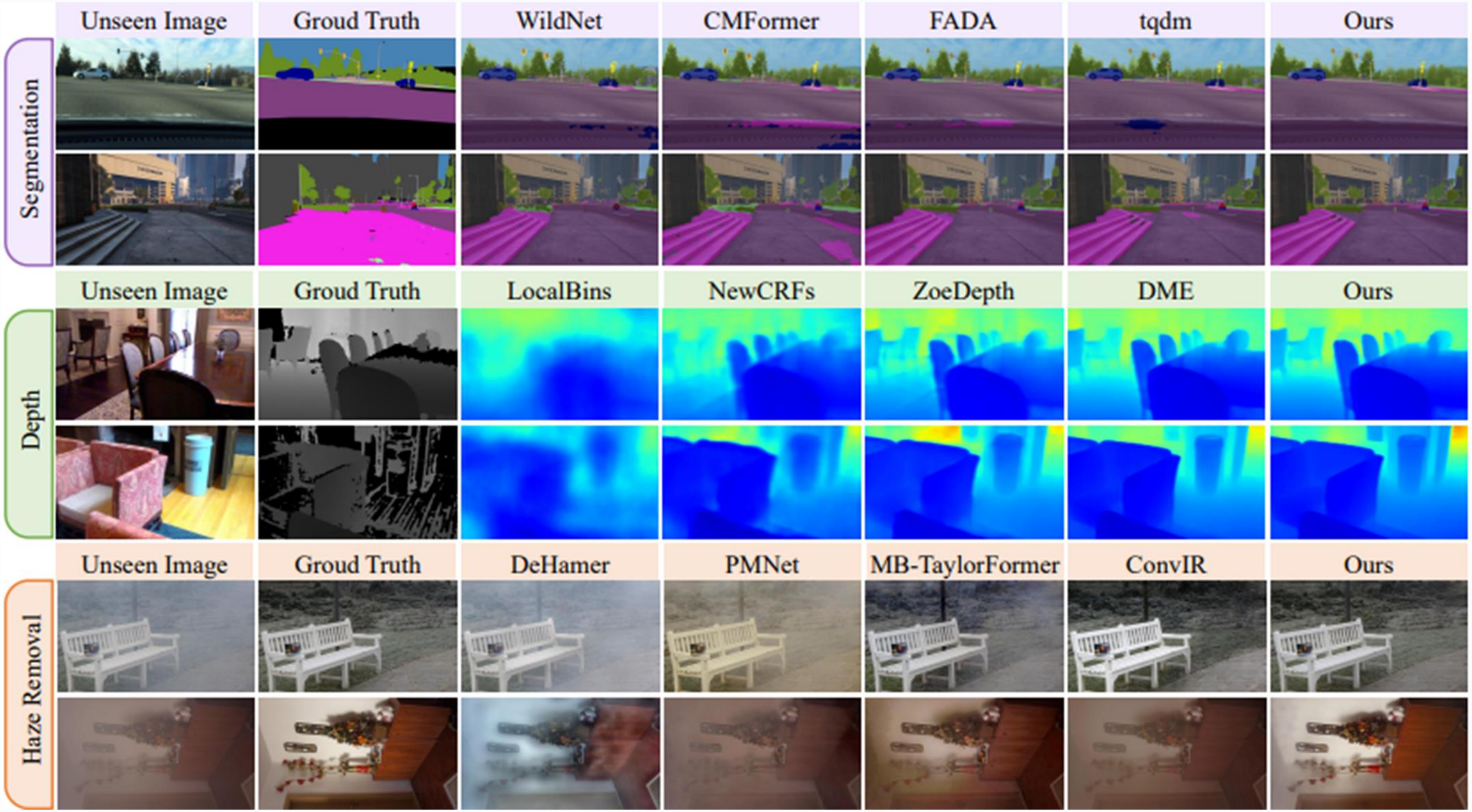
Method	Encoder	Condition	Trained on GTA5 (G) → C	→ B	→ M	Avg.
PTDiffSeg [25]	Diffusion	$\mathcal{I}$+$\mathcal{T}$	52.0	-	-	-
FC-CLIP [87]	ConvNeXt-L	$\mathcal{I}$+$\mathcal{T}$	53.6	47.6	57.4	52.9
DDP [33]*	ConvNeXt-L	$\mathcal{I}$	59.5	56.8	65.7	60.7
DIDEX [53]	Diffusion	$\mathcal{I}$+$\mathcal{T}$	62.0	54.3	63.0	59.7
CLOUDS [2]	ConvNeXt-L	$\mathcal{I}$+$\mathcal{T}$	60.2	57.4	67.0	61.5
DGInStyle [34]	Diffusion	$\mathcal{I}$+$\mathcal{T}$	58.63	52.25	62.47	57.78
HarDiff (Ours)	ConvNeXt-L	$\mathcal{I}$	**62.0**	**58.9**	**68.8**	**63.2**



Method	Virtual KITTI 2 $\delta_1\uparrow$	REL↓	RMSE↓	DIML $\delta_1\uparrow$	REL↓	RMSE↓	DIODE-O $\delta_1\uparrow$	REL↓	RMSE↓	DIODE-I $\delta_1\uparrow$	REL↓	RMSE↓	iBims-1 $\delta_1\uparrow$	REL↓	RMSE↓
Conventional Methods:															
BTS [38] [Arxiv'2019]	0.831	0.115	3.508	0.016	1.785	5.978	0.171	0.837	10.448	0.210	0.418	1.905	0.538	0.231	0.919
AdaBins [3] [CVPR'2021]	0.826	0.123	2.420	0.017	1.941	6.272	0.163	0.663	10.253	0.174	0.443	1.963	0.555	0.212	0.901
LocalBins [4] [ECCV'2022]	0.810	0.127	5.981	0.016	1.820	6.706	0.170	0.821	10.271	0.229	0.412	1.853	0.558	0.211	0.880
NewCRFs [90] [CVPR'2022]	0.829	0.117	2.601	0.199	0.918	6.285	0.173	0.854	9.228	0.187	0.404	1.867	0.548	0.206	0.861
Diffusion Methods:															
DepthGen [64]* [Arxiv'2023]	0.754	0.148	4.632	0.153	2.147	6.873	0.148	0.875	10.362	0.175	0.494	2.030	0.501	0.254	0.932
DDP [33]* [ICCV'2023]	0.862	0.138	2.463	0.249	0.516	2.113	0.468	0.375	5.580	0.605	0.259	1.374	0.574	0.236	0.683
PrimeDepth [93]* [ACCV'2024]	0.829	0.127	2.457	0.159	1.546	5.903	0.167	0.674	10.305	0.209	0.413	1.916	0.529	0.240	0.922
ECoDepth [55] [CVPR'2024]	-	-	-	-	-	-	-	-	-	0.545	0.344	1.164	0.688	0.163	0.664
Marigold [35]* [CVPR'2024]	0.847	0.140	2.529	0.165	1.274	5.072	0.183	0.635	9.046	0.274	0.385	1.547	0.549	0.206	0.895
D4RD [73]* [MM'2024]	0.841	0.138	2.601	0.182	1.305	4.370	0.204	0.627	8.429	0.472	0.409	1.865	0.561	0.218	0.857
DiffusionDepth [22]* [ECCV'2024]	0.868	0.119	2.518	0.257	1.063	2.564	0.296	0.529	8.147	0.529	0.298	1.473	0.596	0.184	0.640
RobustDepth [69]* [ECCV'2024]	0.835	0.102	2.493	0.268	0.924	2.316	0.402	0.397	5.830	0.598	0.235	1.249	0.607	0.187	0.701
Generalization Methods:															
ZoeDepth [5] [Arxiv'2023]	0.850	0.105	5.095	0.292	0.641	3.610	0.208	0.757	7.569	0.386	0.331	1.598	0.615	0.186	0.777
DME [88] [AAAI'2024]	0.840	0.113	4.244	0.199	0.735	3.495	0.215	0.777	9.570	0.479	0.744	0.862	0.585	0.316	0.635
DME-GT [88] [AAAI'2024]	0.881	0.097	3.943	0.296	0.472	2.120	0.508	0.360	5.713	0.654	0.219	0.822	0.589	0.315	0.629
HarDiff (Ours)	**0.901**	**0.092**	**2.406**	**0.301**	**0.458**	**1.997**	**0.524**	**0.318**	**4.957**	**0.667**	**0.215**	0.837	0.638	0.170	**0.620**

Methods	Trained on RESIDE-Indoor SOTS-IN	SOTS-Out	NTIRE-19	NTIRE-20	Trained on RESIDE-Outdoor SOTS-IN	SOTS-Out	NTIRE-19	NTIRE-20
GridDehazeNet [44] [ICCV'2019]	32.14 / 0.98	16.22 / 0.76	09.50 / 0.49	09.01 / 0.40	20.99 / 0.89	29.18 / 0.93	10.16 / 0.50	11.23 / 0.49
DuRN-US [45] [CVPR'2019]	32.12 / 0.98	19.55 / 0.83	10.81 / 0.51	11.27 / 0.51	15.95 / 0.76	19.41 / 0.81	11.04 / 0.51	11.73 / 0.46
FFA-Net [58] [AAAI'2020]	36.36 / 0.98	20.05 / 0.84	10.97 / 0.42	10.70 / 0.44	18.96 / 0.86	30.88 / 0.93	09.64 / 0.50	10.90 / 0.48
DISID [65] [AAAI'2021]	38.91 / 0.98	25.75 / 0.84	16.21 / 0.78	16.28 / 0.67	26.90 / 0.76	30.40 / 0.94	13.36 / 0.52	12.68 / 0.52
DeHamer [27]* [CVPR'2022]	36.92 / 0.97	24.78 / 0.79	12.03 / 0.59	12.71 / 0.60	21.73 / 0.80	27.49 / 0.90	12.91 / 0.51	12.79 / 0.47
PMNet [83]* [ECCV'2022]	36.20 / 0.98	25.92 / 0.84	12.86 / 0.60	12.92 / 0.59	22.82 / 0.81	28.61 / 0.90	13.16 / 0.55	13.04 / 0.51
C^{2}PNet [97]* [CVPR'2023]	37.46 / 0.96	25.41 / 0.83	12.90 / 0.62	13.04 / 0.58	23.09 / 0.82	28.05 / 0.92	13.02 / 0.50	13.81 / 0.55
DDP [33]* [ICCV'2023]	37.39 / 0.97	24.67 / 0.81	14.28 / 0.64	14.16 / 0.62	24.16 / 0.79	29.21 / 0.91	13.36 / 0.51	13.77 / 0.54
MB-TaylorFormer [59]* [ICCV'2023]	36.64 / 0.95	23.83 / 0.80	13.95 / 0.59	14.02 / 0.60	22.08 / 0.78	28.46 / 0.85	12.87 / 0.50	13.50 / 0.52
DiffIR [77]* [ICCV'2023]	37.75 / 0.97	25.19 / 0.84	14.81 / 0.69	14.77 / 0.59	23.93 / 0.77	30.27 / 0.89	13.92 / 0.52	13.98 / 0.53
ConvIR-B [20]* [TPAMI'2024]	38.61 / 0.98	25.88 / 0.85	16.02 / 0.78	16.63 / 0.72	25.72 / 0.81	30.65 / 0.92	14.30 / 0.57	13.96 / 0.55
DCMPNet [95]* [CVPR'2024]	38.90 / 0.98	26.20 / 0.84	15.72 / 0.76	16.00 / 0.60	26.05 / 0.84	31.08 / 0.91	14.04 / 0.58	14.22 / 0.56
DiffLI^{2}D [82]* [ECCV'2024]	39.03 / 0.97	26.07 / 0.85	15.89 / 0.73	16.30 / 0.69	27.90 / 0.83	31.36 / 0.93	14.17 / 0.56	14.25 / 0.53
HarDiff (Ours)	39.76 / 0.99	27.05 / 0.88	17.20 / 0.85	17.61 / 0.75	28.42 / 0.88	32.01 / 0.95	14.99 / 0.61	14.63 / 0.58

Qualitative Results



[12]

Conclusion

• Proposed HarDiff decouples diffusion into low-frequency training and high-frequency sampling, guided by Hartley analysis.

• Achieves state-of-the-art performance on semantic segmentation, depth estimation and dehaze across multiple unseen domains.

• Extensive experiments show the critical role of extreme frequency bands in generalising dense vision models.

• Future work: explore other frequency transforms, combine with adaptive attention, and apply to broader vision tasks.

[13] [14]

Thanks for your attention!

[1]