Frequency-Aware Self-Supervised Monocular Depth Estimation Supplementary Material

Xingyu Chen[1] Thomas H. Li[1,2,3] Ruonan Zhang[1] Ge Li ✉[1]

[1]School of Electronic and Computer Engineering, Peking University [2]Advanced Institute of Information Technology, Peking University

[3]Information Technology R&D Innovation Center of Peking University

cxy@stu.pku.edu.cn tli@aiit.org.cn zhangrn@stu.pku.edu.cn geli@ece.pku.edu.cn

https://github.com/xingyuuchen/freq-aware-depth

1. Pseudo-code of Ambiguity-Masking

The overall algorithm of the proposed Ambiguity-Masking is summarized as Alg. 1. Please refer to our Github repository for full implementation.

Algorithm 1 Extract Ambiguities for Photometric Loss

Input: target image I_t, source images I_{t+n}, indices of source images src_ids, reconstructed images $\tilde{I}_{t+n}$, photometric errors of all source images $\mathcal{L}$

Output: $\mathcal{A}_t^{pe}$: ambiguity mask of the final photometric error

$\mathcal{A}_t \leftarrow$ ExtractAmbiguityForImage(I_t);
$reproj\_ambiguities \leftarrow list$;
for all n in src_ids **do**
  $\mathcal{A}_{t+n} \leftarrow$ ExtractAmbiguityForImage(I_{t+n});
  $\tilde{\mathcal{A}}_{t+n} \leftarrow$ bilinear sample $\mathcal{A}_{t+n}$ subject to $\otimes_{t+n}$; // to get which pixels in reconstructed $\tilde{I}_{t+n}$ are from the ambiguous pixels in source I_{t+n}.
  append $\tilde{\mathcal{A}}_{t+n}$ to $reproj\_ambiguities$;
end for
$min\_idx \leftarrow$ argmin($\mathcal{L}$); // we adopt min. reprojection loss from [12].
$\mathcal{A}'_t \leftarrow reproj\_ambiguities[min\_idx]$; // to gather ambiguity value adopted in the final loss map.
$\mathcal{A}_t^{max} \leftarrow \max(\mathcal{A}_t, \mathcal{A}'_t)$; // as Eq . 13.
$\mathcal{A}_t^{pe} \leftarrow \mathcal{A}_t^{max} < \delta$; // as Eq. 14.
return $\mathcal{A}_t^{pe}$;
procedure ExtractAmbiguityForImage(I)
  $\mathcal{F} \leftarrow$ compute frequency map of I; // as Eq. 9.
  $\mu \leftarrow \nabla_{u+} \cdot \nabla_{u-} < 0 \,||\, \nabla_{v+} \cdot \nabla_{v-} < 0$; // as Eq. 10.
  $\mathcal{A} \leftarrow \mu\mathcal{F}$;
  return $\mathcal{A}$;
end procedure

2. Further Consideration on the Two Modules

We let the Ambiguity-Masking module take input from the Auto-Blur because we want the high-freq regions of input images to be first processed by Auto-Blur before extracting ambiguities. The reason for this lies in the fact that without smoothing the high-frequency areas, the Ambiguity-Masking would wrongly filter out almost all pixels in high-frequency areas as the *dense thin* objects inside are likely to be misjudged as ambiguous colors, disabling them from participate in training.

3. Full Numbers of Hyper-params Ablation

In this section, we show full numbers of ablations of all hyper-parameters in our methods, as reported in Tab. 1. We then give detailed analyses on each one of them.

If δ is too small, the Amb.-masking will wrongly exclude some non-ambiguous pixels, *e.g.*, the long wall from near to far could also satisfy the constraint of gradual color transition, but it does not belong to the problem demonstrated in Fig. 1. If δ is too large, boundaries with little color difference will be missed.

For kernel size s in Auto-Blur, if we decrease s, the receptive field could not be effectively enlarged when measuring pixel similarity. If we increase s too much, the central pixel's contribution (its own characteristic color) is reduced since the Gaussian distribution gets 'shorter' and 'wider'.

For threshold λ, decreasing λ would wrongly smooth the texture-less regions, as the already-weak supervision signal on them will be further weakened. Increasing λ too much would miss some pixels in high-freq regions which could confuse the photometric loss as illustrated in Fig. 2.

For the percentage threshold η of high-frequency pixels in Auto-Blur, when η is too small, not only the texture-less regions but also some object boundary areas which does not belong to 'high-frequency area' would be wrongly smoothed. When η is too large, the same as λ, our Auto-

Hyper-parameter	Value	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
δ	0.2	0.113	0.884	4.814	0.190	0.878	0.960	**0.982**
	0.3	**0.112**	**0.834**	**4.746**	**0.189**	**0.880**	**0.961**	**0.982**
	0.4	0.113	0.864	4.757	0.190	0.879	0.960	**0.982**
s	7	**0.112**	0.836	4.753	0.190	0.878	**0.961**	0.981
	9	**0.112**	**0.834**	**4.746**	**0.189**	**0.880**	**0.961**	**0.982**
	11	0.113	0.868	4.782	**0.189**	0.877	0.960	**0.982**
λ	0.15	0.113	0.844	4.814	0.192	0.879	0.959	0.982
	0.20	**0.112**	**0.834**	**4.746**	**0.189**	**0.880**	**0.961**	**0.982**
	0.25	0.113	0.881	4.797	0.191	0.877	0.959	0.981
η	50	0.113	0.860	4.804	0.192	0.875	0.959	0.981
	60	**0.112**	**0.834**	**4.746**	**0.189**	**0.880**	**0.961**	**0.982**
	70	0.114	0.887	4.839	0.190	0.878	0.960	**0.982**

Table 1. Ablations on all hyper-parameters.

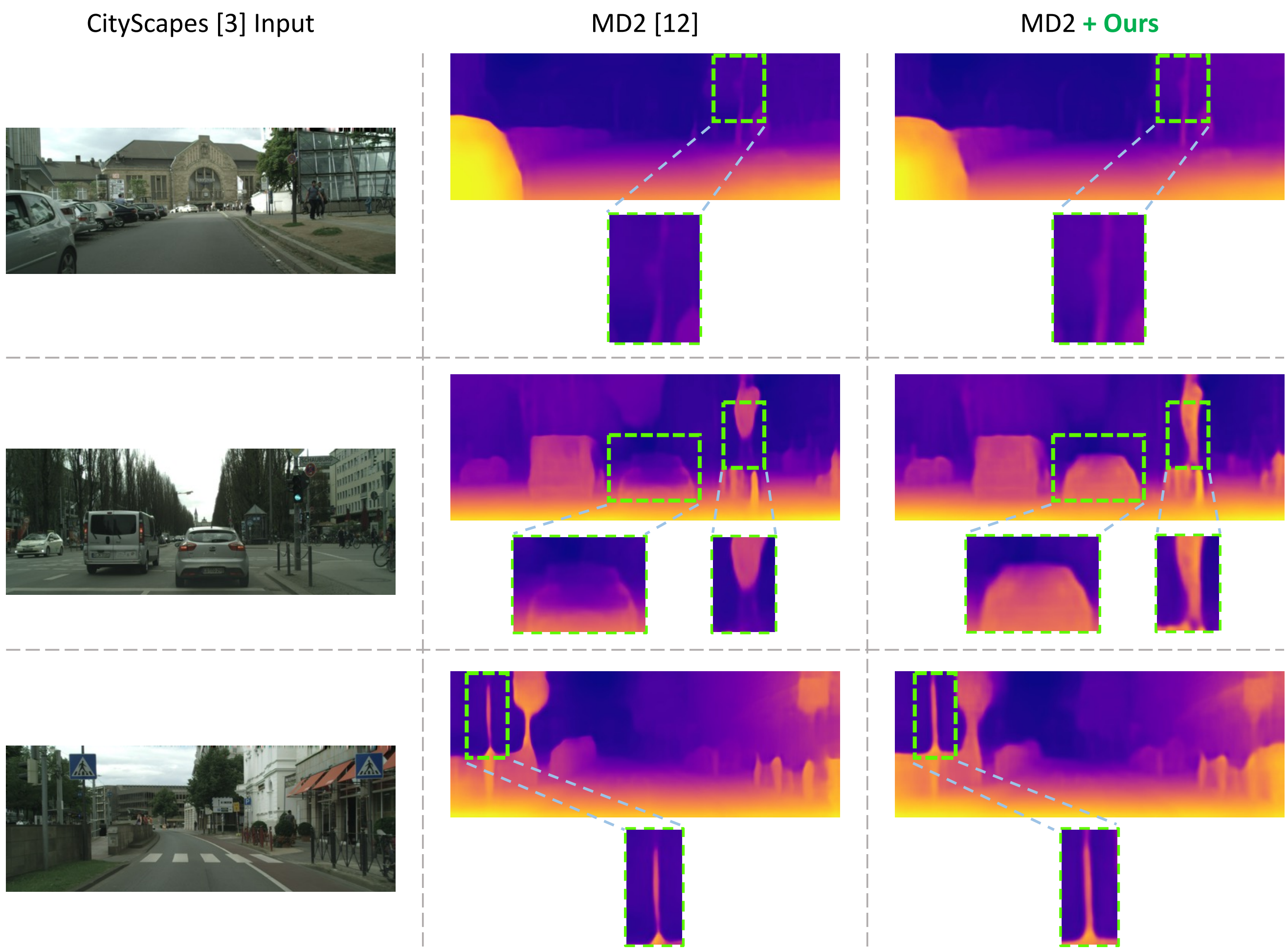

Figure 1. High-resolution qualitative comparisons of *Monodepth2* [12] with and w/o our proposed methods (input from CityScapes [3]).

Blur would be too strict, *i.e.* miss to smooth some pixels in high-frequency areas which could confuse the photometric loss.

4. Full-Resolution Qualitative Results

We show more full-resolution qualitative depth predictions in Fig. 1 (CityScapes) and Fig. 2 (KITTI).

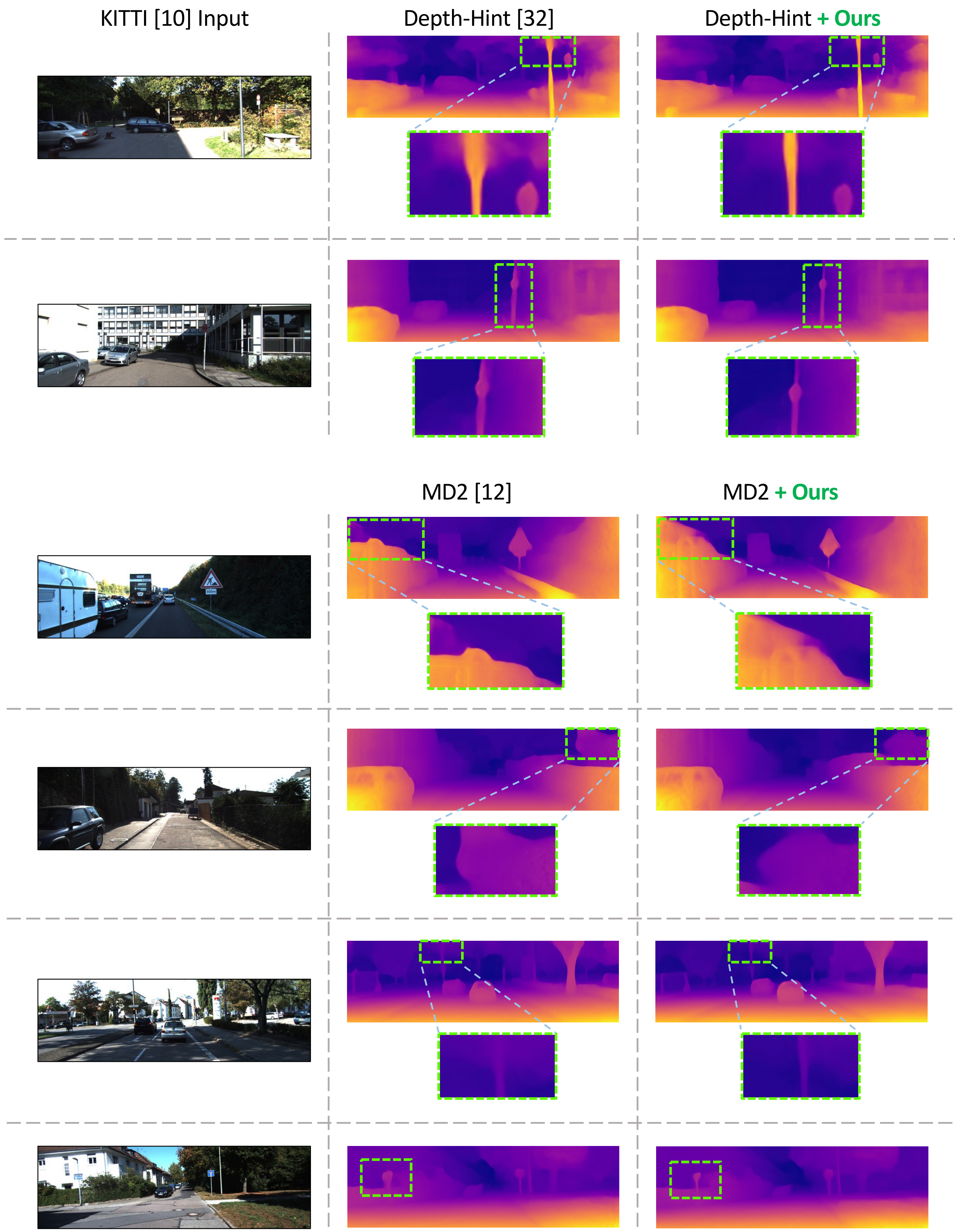

Figure 2. High-resolution qualitative comparisons of *Depth-Hints* [32] and *Monodepth2* [12] with and w/o our proposed methods (input from KITTI [10]).