Rethinking Depth Estimation for Multi-view Stereo: A Unified Representation (Supplementary Material)

Rui Peng [1] Rongjie Wang [2] Zhenyu Wang [1] Yawen Lai [1] Ronggang Wang ✉[1,2]
[1]School of Electronic and Computer Engineering, Peking University [2]Peng Cheng Laboratory
ruipeng@stu.pku.edu.cn rgwang@pkusz.edu.cn
https://github.com/prstrive/UniMVSNet

Abstract

Depth estimation is solved as a regression or classification problem in existing learning-based multi-view stereo methods. Although these two representations have recently demonstrated their excellent performance, they still have apparent shortcomings, e.g.*, regression methods tend to overfit due to the indirect learning cost volume, and classification methods cannot directly infer the exact depth due to its discrete prediction. In this paper, we propose a novel representation, termed* ***Unification****, to unify the advantages of regression and classification. It can directly constrain the cost volume like classification methods, but also realize the sub-pixel depth prediction like regression methods. To excavate the potential of unification, we design a new loss function named* ***Unified Focal Loss****, which is more uniform and reasonable to combat the challenge of sample imbalance. Combining these two unburdened modules, we present a coarse-to-fine framework, that we call* ***UniMVSNet****. The results of ranking* ***first*** *on both DTU and Tanks and Temples benchmarks verify that our model not only performs the best but also has the best generalization ability.*

A. More Explanation of Unified Focal Loss

As shown in Equation (9), the dedicated function to control the range of scaling factor is designed as the sigmoid-like function as:

$$S_b(x) = \frac{1}{(1 + b^{-x})} \tag{a}$$

where $x = \frac{|q-u|}{q^+}$ in this paper and its range is $[0, +\infty)$, therefore, the range of $S_b(x)$ is $[0.5, 1)$. As aforementioned, we adopt an asymmetrical scaling strategy to protect the precious positive learning signals and scale the range of S_5^+ to $[1, 3)$ and S_5^- to $[0, 1)$. The detailed implementation of S_5^+ is:

$$S_5^+(x) = 4 \times (\frac{1}{1 + 5^{-x}} - 0.5) + 1 \tag{b}$$

and the detailed implementation of S_5^- is:

$$S_5^-(x) = 2 \times (\frac{1}{1 + 5^{-x}} - 0.5) \tag{c}$$

B. Finer DTU Ground-truth

As mentioned in our main paper, we adopt the finer ground-truth to train our model additionally for a fair comparison with the start-of-the-art methods [3]. The refinement of each DTU ground-truth is achieved by cross-filtering with its neighbor viewpoints. For convenience, we directly adopt the processed results provided in [3], and we only adopt the mask which indicates the validity of each point. Concretely, we adopt the union of the mask provided in [1,4] and the up-sampled mask provided in [3] as the final mask.

C. More Ablation Studies on DTU Dataset

Here, we perform more ablation studies on DTU to show you more information about our implementation.

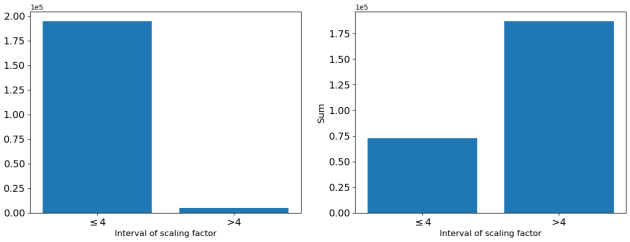

Figure A. **The statistics of scaling factor $x = \frac{|q-u|}{q^+}$ in Eq. (8).**

The scaling factor in UFL. The range of scaling factor in Eq. (8) is $[0, +\infty)$. We count the average number and sum of scaling factors that fall in different intervals. As shown in Fig. A, most of the scaling factors are less than 4 (Left figure). Even though, those small fractions of larger scaling factors take more weight (Right figure). This results in those abnormally large scaling factors occupying

the model's training, and lead to difficulty in model convergence and extremely poor performance. Therefore, we introduce a dedicated function to control the scaling factors' range.

The dedicated function in UFL. As described in our main paper, we design the positive dedicated function as Eq. (b) and negative dedicated function as Eq. (c). In fact, this final implementation is confirmed under our more experimental results. As shown in Tab. A, compared to adopting a common *sigmoid* function with a base e, it's better to use a dedicated function with a base 5. Meanwhile, scaling the range of dedicated function to $[1, 3)$ is better than $[1, 2)$. As shown in Fig. B, the base number controls the speed at which the function converges to the maximum value. The smaller the base number, the slower the convergence. In our experiments, we found that most of the points whose $x = \frac{|q-u|}{q^+}$ is in the interval $[0, 4]$, so the scaling value calculated by the dedicated function needs to be distinguishable for the points in this interval, and we set $b = 5$ in this paper. To be honest, we have only conducted a limited number of the base number the range of dedicated function as shown in Tab. A due to the time and resource considerations, and we believe there will be more powerful configurations.

Base Number	Range	ACC.(mm)	Comp.(mm)	Overall(mm)
e	$[1, 2)$	0.354	0.282	0.318
5	$[1, 2)$	0.354	0.280	0.317
5	$[1, 3)$	0.352	0.278	0.315

Table A. **Ablation results of dedicated function.** While the base number is ablated for both the positive and negative dedicated function, we only ablate the range of positive dedicated function.

The tunable parameter in UFL. Tunable parameters in UFL like α and λ are also important for rebalancing samples. As mentioned in our main paper, we always set $\alpha^+ = 1$ to protect the positive learning signals and configure other tunable parameters stage by stage due to the different number of depth hypotheses. In our implementation, we set the number of depth hypotheses to 48, 32, and 8 from stage1 to stage3. Apparently, the sample imbalance in stage1 is the most challenging, while it's the most relaxing or even negligible in stage3. As shown in Tab. B, applying the same configuration across all stages performs the worst which indicates that the imbalances faced by different stages are different.

Proximity $\mathcal{VS}$. Offset. Different from the *Regression* and *Classification*, we propose *Unification* to classify the optimal depth hypothesis and regress its offset to ground-truth depth simultaneously. As shown in Fig. C, there are two ways to regress the offset. The first is to predict proximity which is the complement of the offset and is also the method we adopt in this paper. The second is to directly estimate the offset. The comparison results between them are shown in Tab. C. We can see that adopting proximity to regress the offset indirectly is much more powerful than purely using offset. As mentioned in our main paper, our *Unification* can be decomposed into two parts: classify the optimal depth hypothesis first and then regress the proximity for it. We infer that the reason why proximity is better than offset is the positive relationship between the magnitude of proximity and the quality of the classified optimal depth hypothesis in the first step. Meanwhile, we think that there should be better settings to improve the performance of using offset, but we have not made more attempts here.

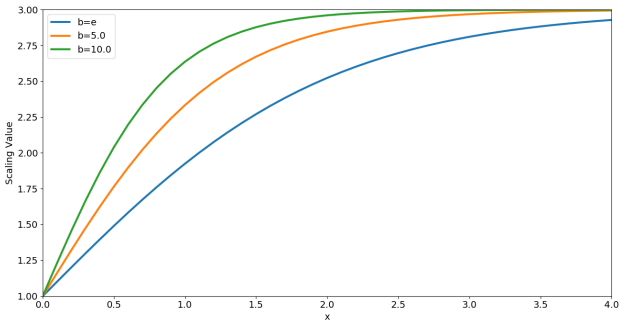

Figure B. **The scaling value obtained through the positive dedicated function under different base numbers.**

α^-			λ			ACC.(mm)	Comp.(mm)	Overall(mm)
stage1	stage2	stage3	stage1	stage2	stage3			
0.75	0.75	0.75	2	2	2	0.347	0.325	0.336
1	1	1	2	1	0	0.366	0.282	0.324
0.75	0.50	0.25	2	1	0	0.353	0.287	0.320

Table B. **Ablation results of tunable parameters.** These experiments are conducted on the model with only our *Unification* and UFL, and don't adopt the finer DTU ground-truth or adaptive aggregation.

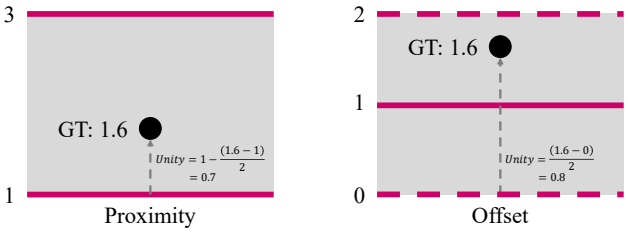

Figure C. **Two available solutions to generate *Unity* (supervised signal) from the ground-truth depth.** For proximity, we use the interval above the depth hypothesis as its regression interval, and for offset, use the area where the depth hypothesis is the median value as the regression interval.

Method	ACC.(mm)	Comp.(mm)	Overall(mm)
Offset	0.429	0.336	0.383
Proximity	0.372	0.282	0.327

Table C. **Comparison of proximity and offset.** These experiments are conducted on the model only using our *Unification*.

D. More Comparisons between Unification, Classification and Regression

(1) Regression methods are harder to converge and have a greater risk of overfitting due to its indirect learning strategy, which has been studied in [5]. Meanwhile, they tend to generate smooth depth in object boundaries, because they treat the depth as the expectation of the depth hypotheses. However, they can achieve sub-pixel depth estimation, therefore, they have better accuracy. **(2)** Classification methods cannot generate accurate depth due to their discrete prediction, but they constrain the cost volume directly and achieve better completeness. **(3)** Our unification is exactly the complement of these two approaches. *Take the essence, get rid of the dross*. On the one hand, We directly constrain the cost volume to keep the model robust and pick the regression interval with the maximum unity to maintain the sharpness of the object boundary. On the other hand, we regress the proximity in the picked regression interval to generate accurate depth. Therefore, our unification is hoped to achieve regression's accuracy and classification's completeness. The results shown in Tab. 3 just prove this.

E. Limitation

As aforementioned, there are several tunable parameters in our Unified Focal Loss that will affect performance. The process of finding a satisfactory parameter configuration is a cumbersome challenge for newcomers. On the other hand, as long as we have a sufficient understanding of Focal Loss [2], this process will become handy. Anyway, an adaptive form or a form with fewer parameters will be a more concise and efficient choice.

F. More Results on DTU Dataset

Figure D shows our additional point clouds reconstruction results. It can be seen that the point cloud reconstructed by our method has excellent accuracy and completeness.

References

[1] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, pages 2495–2504, 2020. 1

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 3

[3] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *ICCV*, 2021. 1

[4] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, pages 767–783, 2018. 1

[5] Youmin Zhang, Yimin Chen, Xiao Bai, Suihanjin Yu, Kun Yu, Zhiwei Li, and Kuiyuan Yang. Adaptive unimodal cost volume filtering for deep stereo matching. In *AAAI*, pages 12926–12934, 2020. 3

Figure D. **More qualitative results on DTU dataset.**