Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching

신 민 정

Vision & Display Systems Lab.

Dept. of Electronic Engineering, Sogang University

Outline

- Introduction
  - Stream of Depth Estimation and 3D-reconstruction
- Background
  - Stereo image and Multi-View Stereo
  - Stereo Matching and disparity & Rectification
- Related work
  - MVSNet: Depth Inference for Unstructured Multi-view Stereo
- Paper_Cascaded MVS
  - Contributions
  - Methodology
  - Experiment Results
- Conclusion
  - Results and Limits
  - Reference

Introduction

- Stream of Depth Estimation and 3D reconstruction

1. Camera parameter & pose

- Structure-from–motion revisited. 2016. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)

2. Depth map

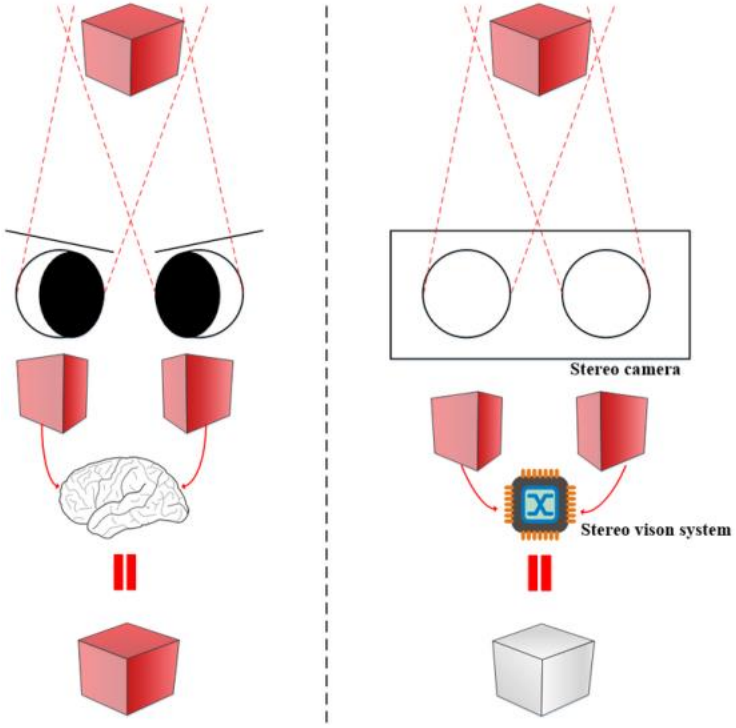
- Extreme View Synthesis.2019.IEEE/CVF International Conference on Computer Vision(ICCV)
- DPSNet:End-to-end Deep Plane Sweep Stereo.2019.Computer Science

3. Both

- DeepSFM: Structure from motion via deep bundle adjustment. published in ECCV 2020

Background

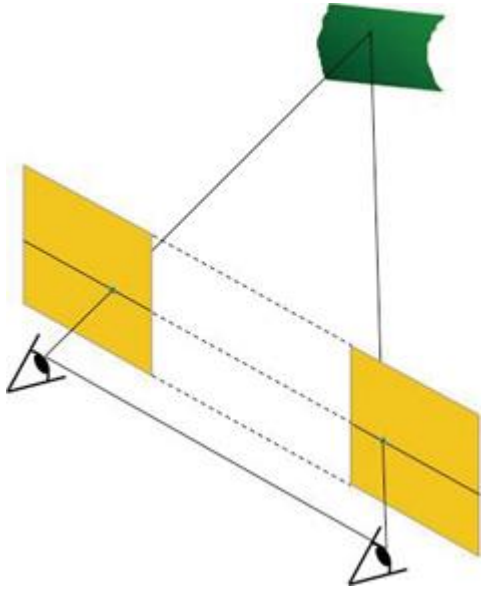
1. Stereo Image

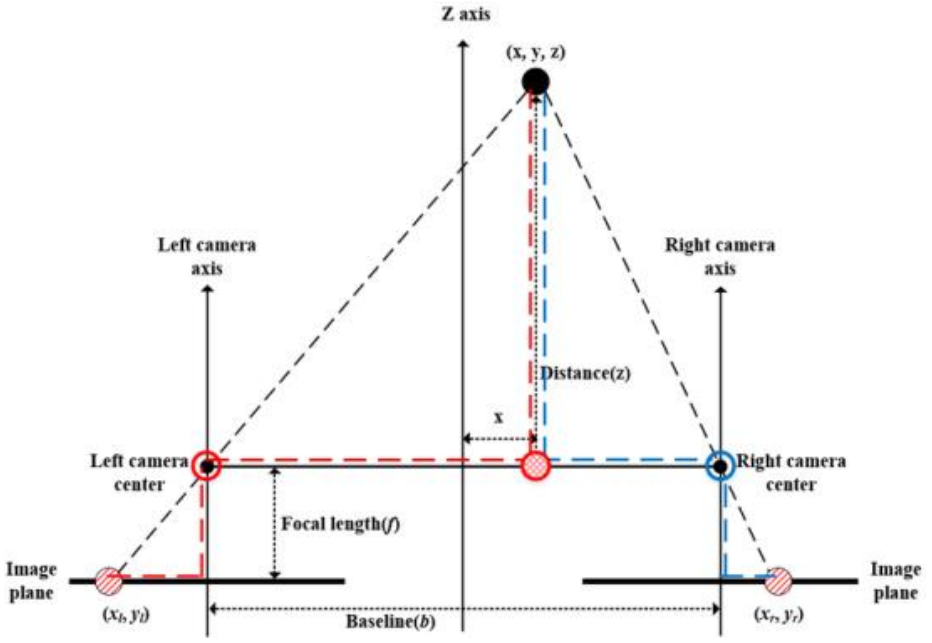


Different photos taken from different locations are required to restore missing distance information

Background

- Disparity(시차)



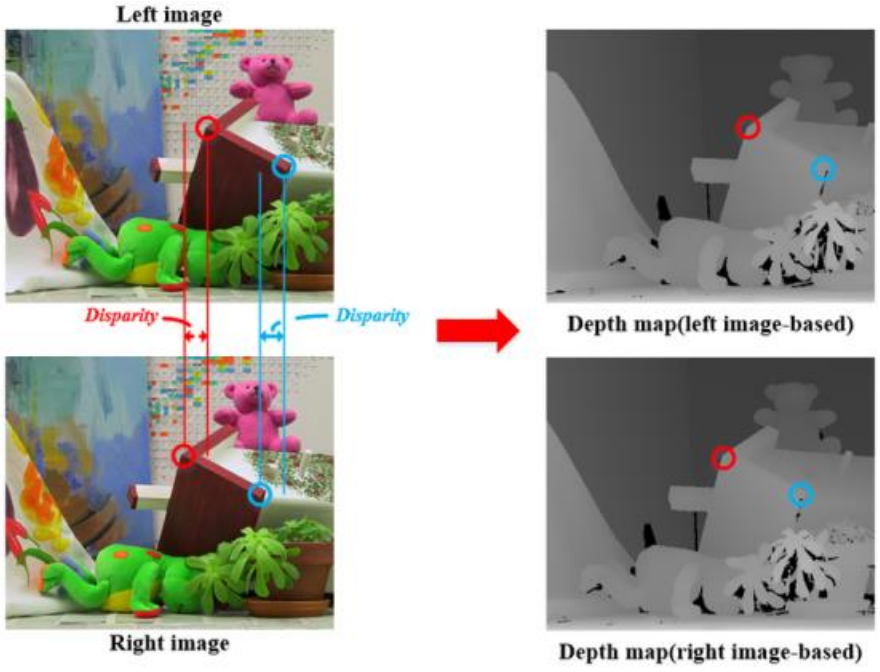


$$disparity = x - x' = \frac{B \cdot f}{z}$$

The difference in position on the x-axis
Is called 'disparity'

Background

2. Stereo matching



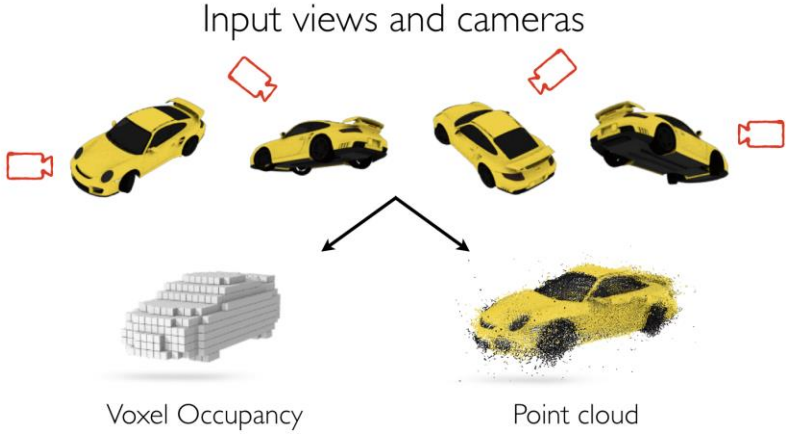
By stereo matching, disparity is computed.

3. Multi-View Stereo

Stereo matching (2-view)

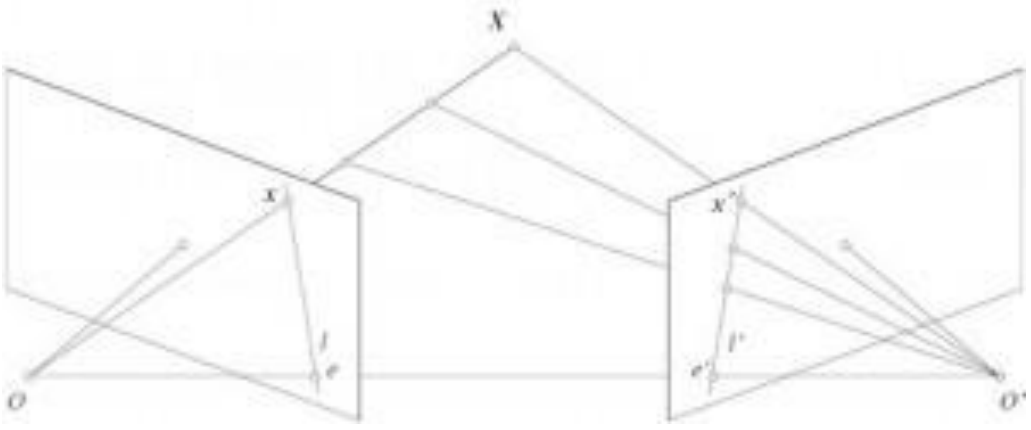
Input >= 3 images and camera poses

Output: Depth map to point cloud



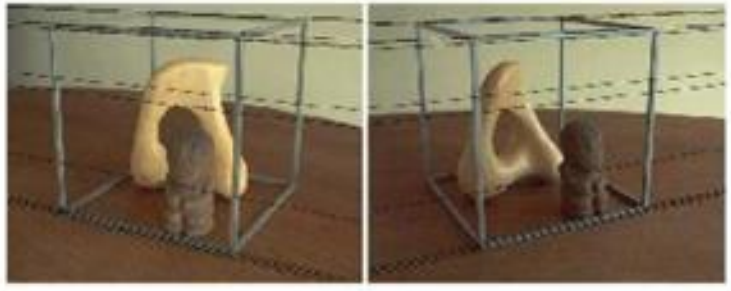
Background

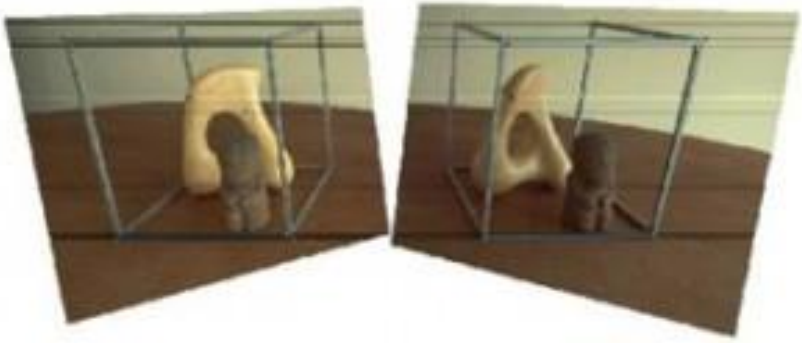
4. Rectification



Matching is too time consuming
So, Use Epipolar constraint

Matching epipolar lines parallel is called rectification





After Rectification, start stereo matching

It Matches y-axis and compute disparity

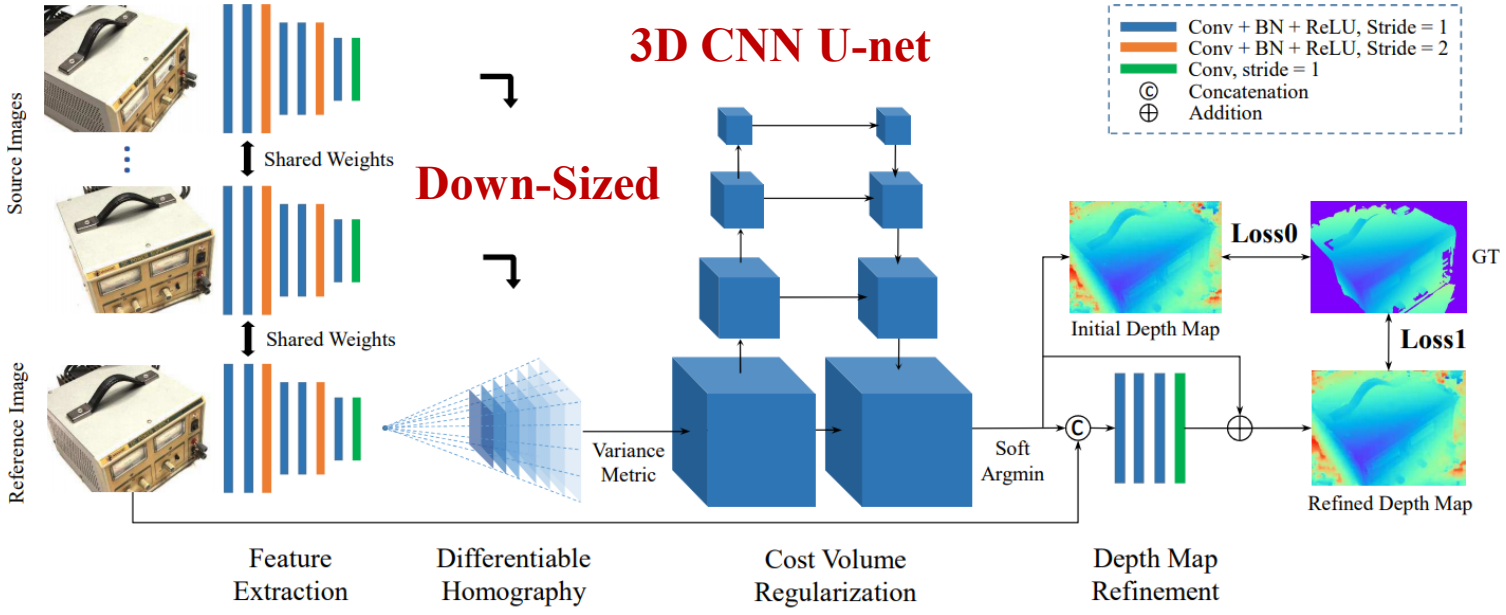
Related Work

Variance Cost Metric

$$\mathbf{C} = \mathcal{M}(\mathbf{V}_1, \cdots, \mathbf{V}_N) = \frac{\sum_{i=1}^{N} (\mathbf{V}_i - \overline{\mathbf{V}_i})^2}{N}$$

Prior research: MVSNet

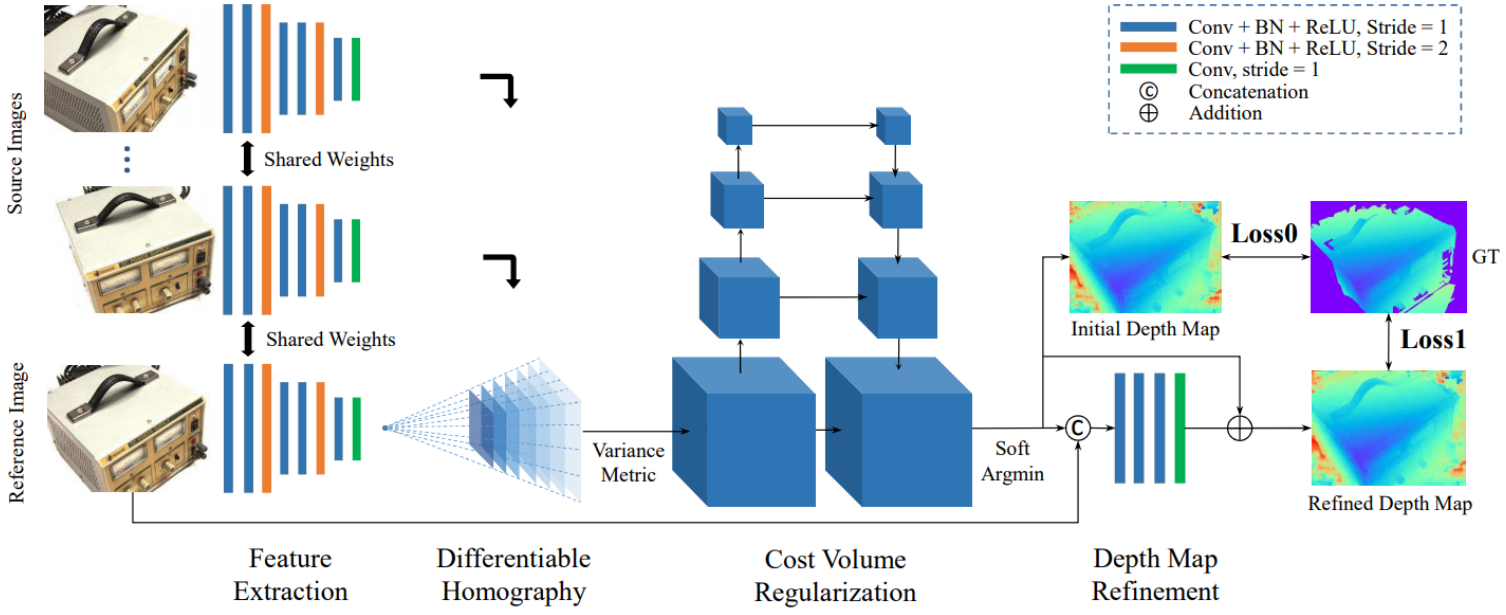
Camera parameter is needed, output: Depth map



Flexibly adapts arbitrary N-view inputs using a variance-based cost metric that maps multiple features into one cost feature.

Related Work

$$Loss = \sum_{p \in \mathbf{p}_{valid}} \underbrace{\|d(p) - \hat{d}_i(p)\|_1}_{Loss0} + \lambda \cdot \underbrace{\|d(p) - \hat{d}_r(p)\|_1}_{Loss1}$$



Softmax operation with 1-channel volume for probability normalization

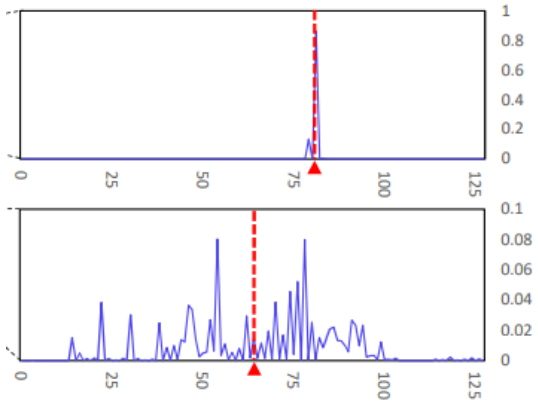
Softmax operation with 1-channel volume for probability normalization

Related Work

1. Initial Depth map

$$\mathbf{D} = \sum_{d=d_{min}}^{d_{max}} d \times \mathbf{P}(d)$$

(d) Probability Map



(c) Probability distribution

2. Probability map

Based on the initial depth map calculated above, the probability sum of the four closest virtual planes is calculated.

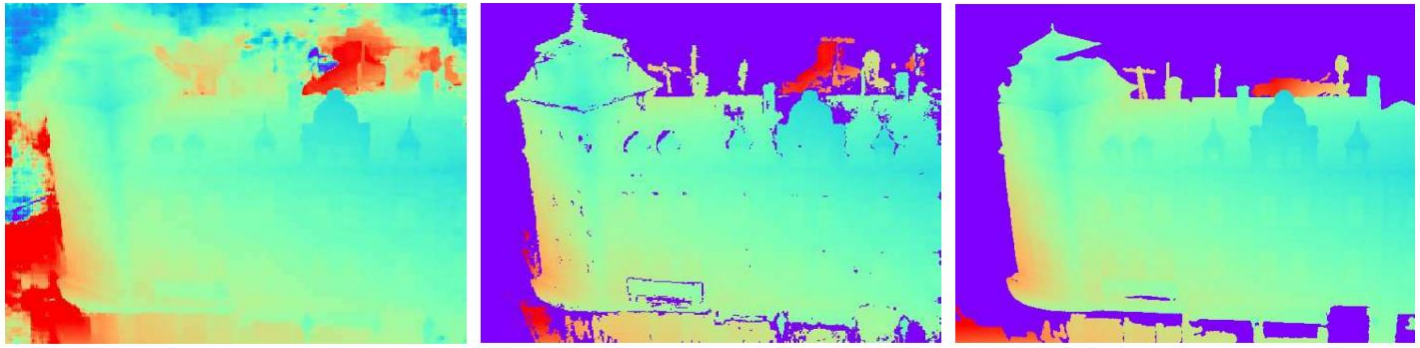
After that, thresholding is performed and outlier filtering is performed.

3. Depth map Refinement

To restore the boundary information well, they apply a depth residual learning network at the end of MVSNet.

Related Work

4. Depth map Filtering



(a) Inferred depth map (b) Filtered depth map (c) GT depth map

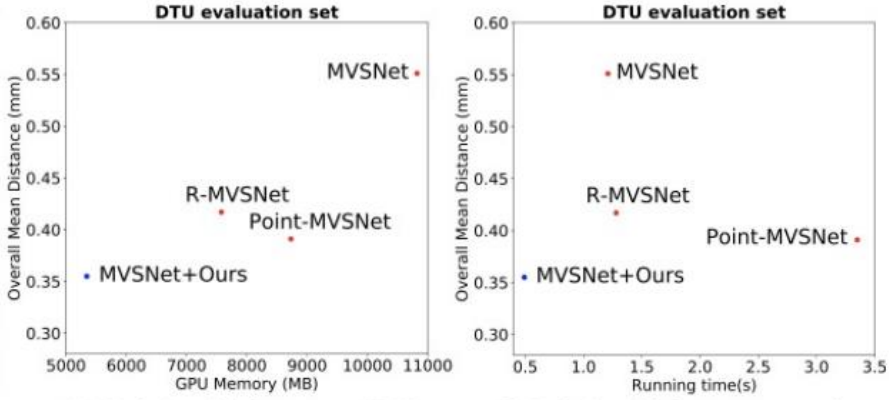
5. Depth map Fusion

To integrate depth maps from different views to a unified point cloud representation.

Paper_Cascaded MVS

Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching.2020 CVPR

- **Contribution**
  - MVSNet은 low-resolution이라는 제한
  - 3D cost volume을 구할 때 메모리와 연산시간이 많이 들기 때문에 high-resolution Output은 무리
  - 이와 같은 문제를 해결
  - 3D reconstruction 정확도도 높이는 동시에 연산속도 및 메모리 효율성까지 잡음



(e) Relationship between GPU memory and accuracy

(f) Relationship between running time and accuracy

Paper_Cascaded MVS

- **Methodology**

Using feature pyramid network to extract multi-scale features

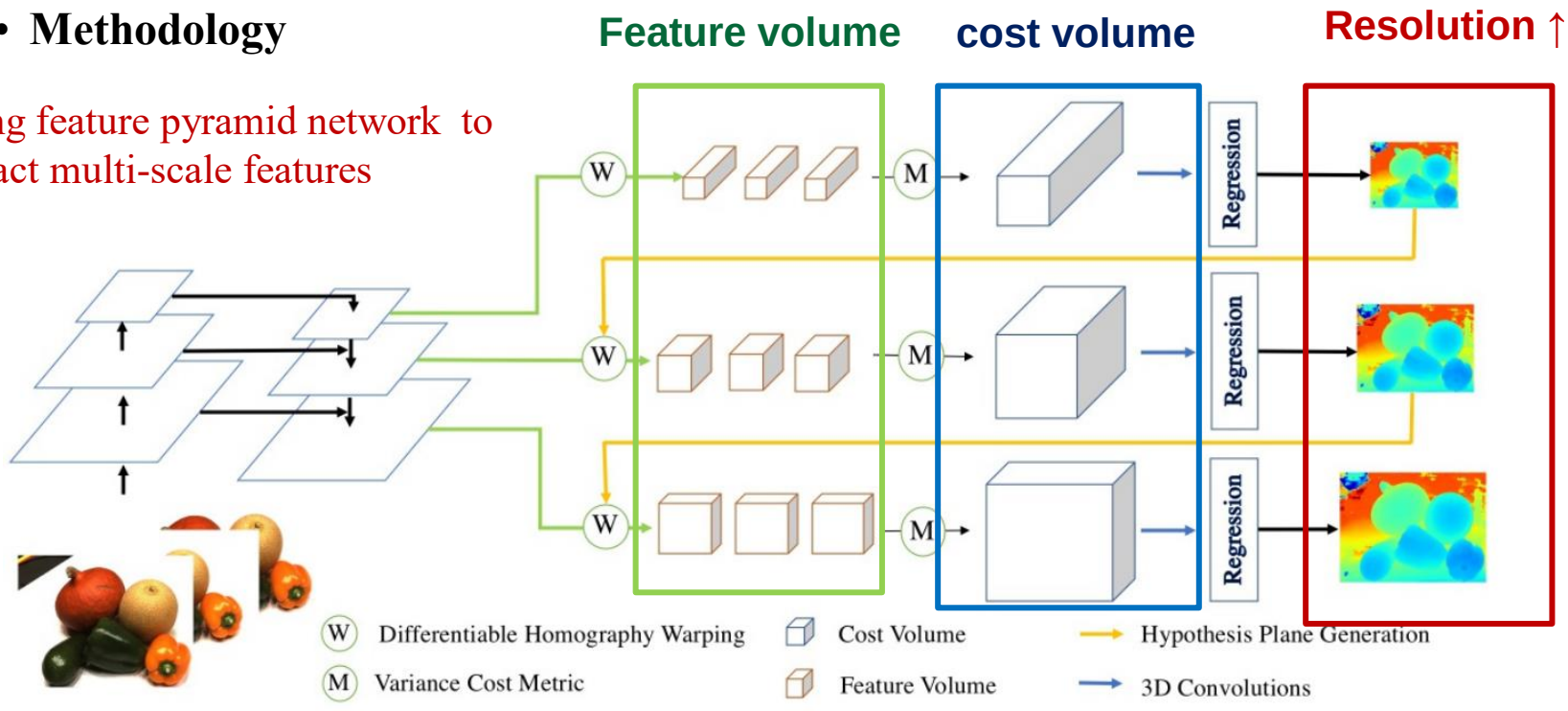


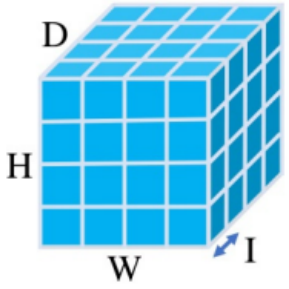
Figure 2: Network architecture of the proposed cascade cost volume on MVSNet [52], denoted as MVSNet+Ours.

Paper_Cascaded MVS

- **Methodology**

Multi-View Stereo

1. Back bone: MVSNet
2. Cost Volume Formulation
3. Camera parameter is needed
4. Output : Depth to point cloud



	Plane Num.	Plane Interv.	Spatial Res.
Efficiency	Negative	Positive	Negative
Accuracy	Positive	Negative	Positive

Figure 3: **Left**: the standard cost volume. D is the number of hypothesis planes, W × H is the spatial resolution and I is the plane interval. **Right**: The influence factors of efficiency (run-time and GPU memory) and accuracy.

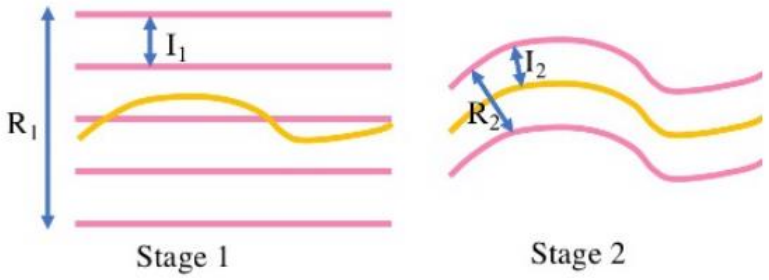
Paper_Cascaded MVS

- **Methodology**

1. Cascaded- cost Volume

- H×W ×D×F = 1600×1184×256×32

- To resolve the problems above,
cascade cost volume formulation is needed



A. Hypothesis Range

$$R_{k+1} = R_k \cdot w_k$$

$w_k < 1$ Reducing factor of hypothesis range

B. Hypothesis Plane Interval

$$I_{k+1} = I_k \cdot p_k,$$

$p_k < 1$ Reducing factor of hypothesis plane Interval

C. Number of Hypothesis Planes

$$D_k = R_k / I_k$$

D. Spatial Resolution

$$\frac{W}{2^{N-k}} \times \frac{H}{2^{N-k}}$$

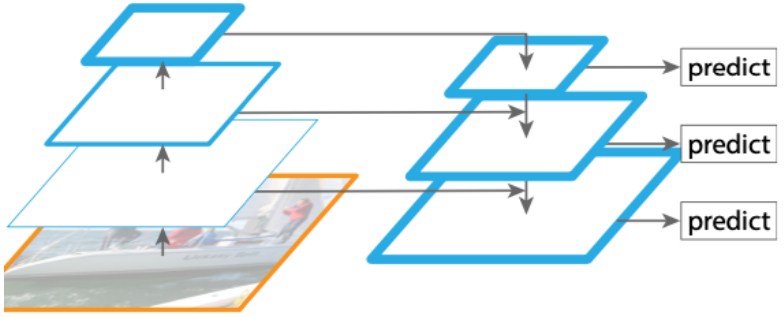
N: total stage number of cascade cost volume

Paper_Cascaded MVS

- **Methodology**

2. Feature pyramid

The standard cost volume is constructed using the top level feature maps which contains **high-level semantic features** but **lacks low-level** finer representations.



(d) Feature Pyramid Network

3. Loss Function

$$Loss = \sum_{k=1}^{N} \lambda^k \cdot L^k$$

Paper_Cascaded MVS

- **Methodology**

Stereo Matching

Backbone: PSMNet

$$C_r(d) = X_l - d$$

It is already rectified.

To build feature volumes,
Warp the feature maps of the right view to the left view using only **translation** along the **x-axis**

Multi-View Stereo

Backbone: MVSNet

$$H_i(d) = K_i \cdot R_i \cdot (I - \frac{(t_1 - t_i) \cdot {n_1}^T}{d}) \cdot {R_1}^T \cdot {K_1}^{-1}$$

Cascaded redefining

$$H_i(d_k^m + \Delta_{k+1}^m) = K_i \cdot R_i \cdot (I - \frac{(t_1 - t_i) \cdot {n_1}^T}{d_k^m + \Delta_{k+1}^m}) \cdot {R_1}^T \cdot {K_1}^{-1}$$

$$C_r(d_k^m + \Delta_{k+1}^m) = X_l - (d_k^m + \Delta_{k+1}^m)$$

Paper_Cascaded MVS

- **Experiment Results_ablation**

	Depth Num.	Depth Interv.	Acc.	Comp.	Overall
MVSNet	192	1	0.4560	0.6460	0.5510
MVSNet-Cas$_2$	96, 96	2, 1	**0.4352**	0.4275	0.4314
MVSNet-Cas$_3$	96, 48, 48	2, 2, 1	0.4479	**0.4141**	**0.4310**
MVSNet-Cas$_4$	96, 48, 24, 24	2, 2, 2, 1	0.4354	0.4374	0.4364
MVSNet-Cas$_3$-share	96, 48, 48	2, 2, 1	0.4741	0.4282	0.4512

Stages	Resosution	>2mm(%)	>8mm(%)	Overall (mm)	GPU Mem. (MB)	Run-time (s)
1	1/4 × 1/4	0.310	0.163	0.602	2373	0.081
2	1/2 × 1/2	0.208	0.084	0.401	4093	0.243
3	1	0.174	0.077	0.355	5345	0.492

Table 3: The statistical results of different stages in cascade cost volume. The statistics are collected on the DTU evaluation set [1] using MVSNet+Ours. The run-time is the sum of the current and previous stages. The base of resolution of input images in this experiment is 1152 × 864.

Reconstruction accuracy with cas3 is the best
So N=3

Not sharing cascaded stage parameter is better

Paper_Cascaded MVS

- **Experiment Results_ablation**

Methods	cascade?	upsample?	feature pyramid?	Acc. (mm)	Comp. (mm)	Overall (mm)
MVSNet	×	×	×	0.456	0.646	0.551
MVSNet-Cas$_3$	✓	×	×	0.450	0.455	0.453
MVSNet-Cas$_3$-Ups	✓	✓	×	0.419	0.338	0.379
MVSNet+Ours	✓	×	✓	0.325	0.385	0.355

Cascaded MVS leads better results than just bilinear up-sampling feature map

Paper_Cascaded MVS

- **Experiment Results_multi-view stereo**

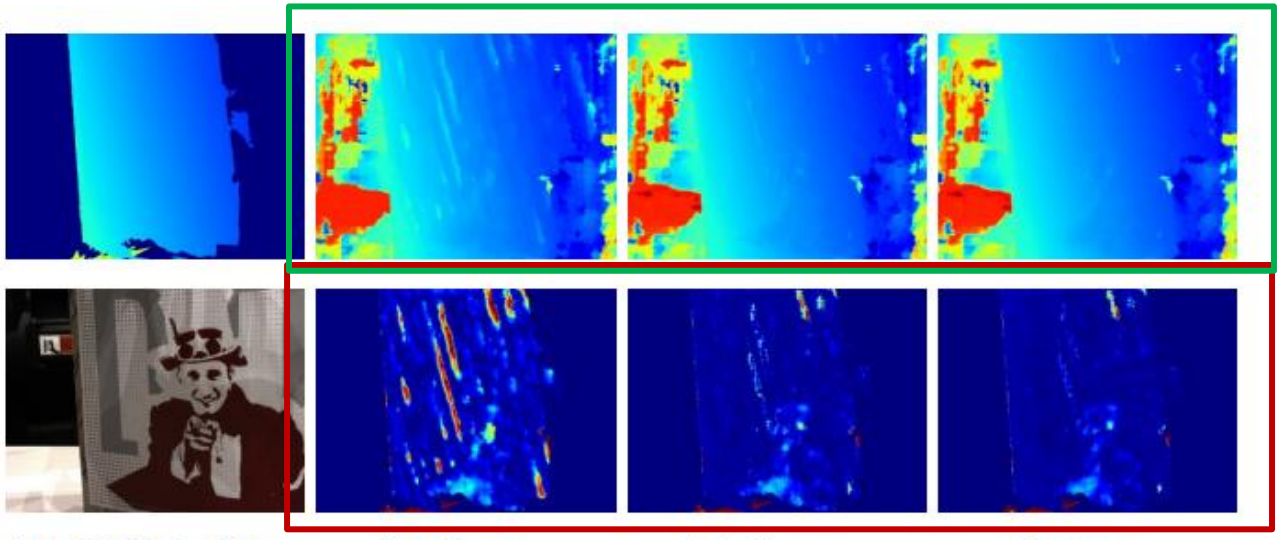
	Rank	Mean	Family	Francis	Horse	Lighthouse	M60	Panther	Playground	Train
COLMAP [39, 40]	54.62	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04
R-MVSNet [53]	40.12	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
Point-MVSNet [4]	38.12	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06
ACMH [47]	15.00	54.82	69.99	49.45	45.12	59.04	52.64	52.37	58.34	51.61
P-MVSNet [29]	12.25	55.62	70.04	44.64	40.22	**65.20**	55.08	**55.17**	**60.37**	**54.29**
MVSNet [52]	52.00	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
MVSNet+Ours	**9.50**	**56.42**	**76.36**	**58.45**	**46.20**	55.53	**56.11**	54.02	58.17	46.56

Statistical& reconstruction results on the Tanks and Temples dataset of state-of-the-art multi-view stereo and our methods

Figure 6: Point cloud results of MVSNet+Ours on the intermediate set of Tanks and Temples dataset [24].

Paper_Cascaded MVS

- **Experiment Results_multi-view stereo**



Depth map

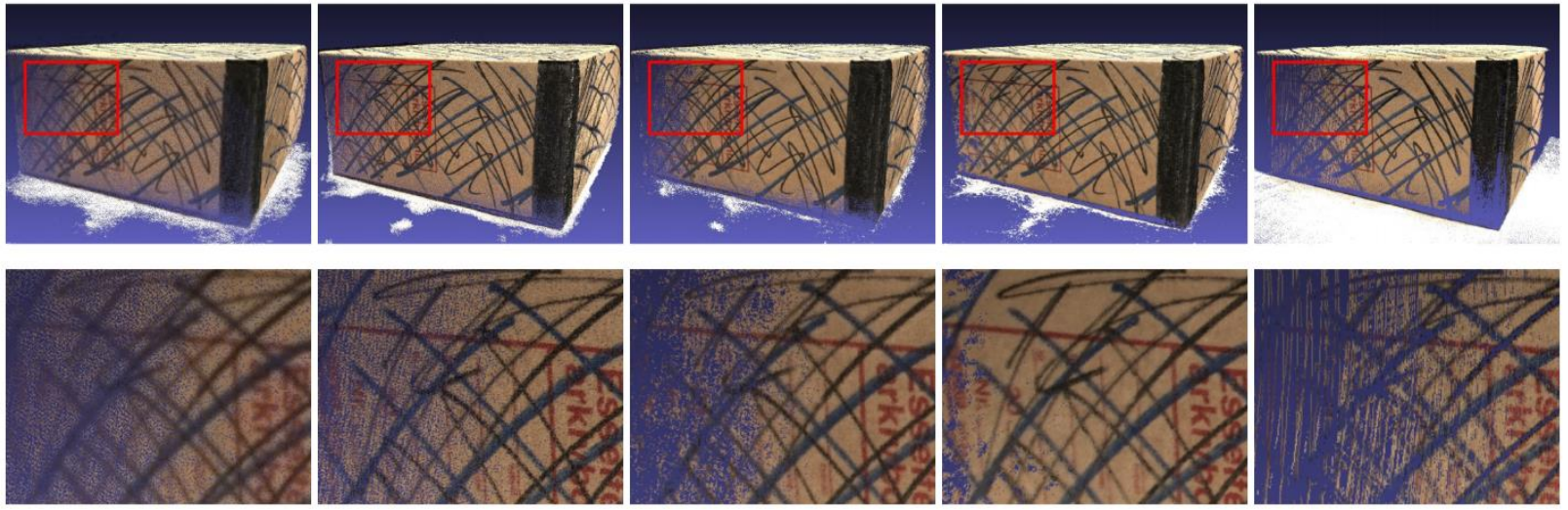
Error map

(a) GT&Ref Img (b) Stage$_1$ (c) Stage$_2$ (d) Stage$_3$

Figure 7: Reconstruction results of each stage. **Top row**: Ground truth depth map and intermediate reconstructions. **Bottom row**: Error maps of intermediate reconstructions.

Paper_Cascaded MVS

- **Experiment Results_multi-view stereo**



(a) MVSNet [52] (b) R-MVSNet [53] (c) Point MVSNet [4] (d) MVSNet+Ours (e) Ground Truth

Figure 5: Multi-view stereo qualitative results of scan 10 on DTU dataset [1]. **Top row**: Generated point clouds of different methods and ground truth point clouds. **Bottom row**: Zoomed local areas.

Paper_Cascaded MVS

- **Experiment Results_Stereo Matching**

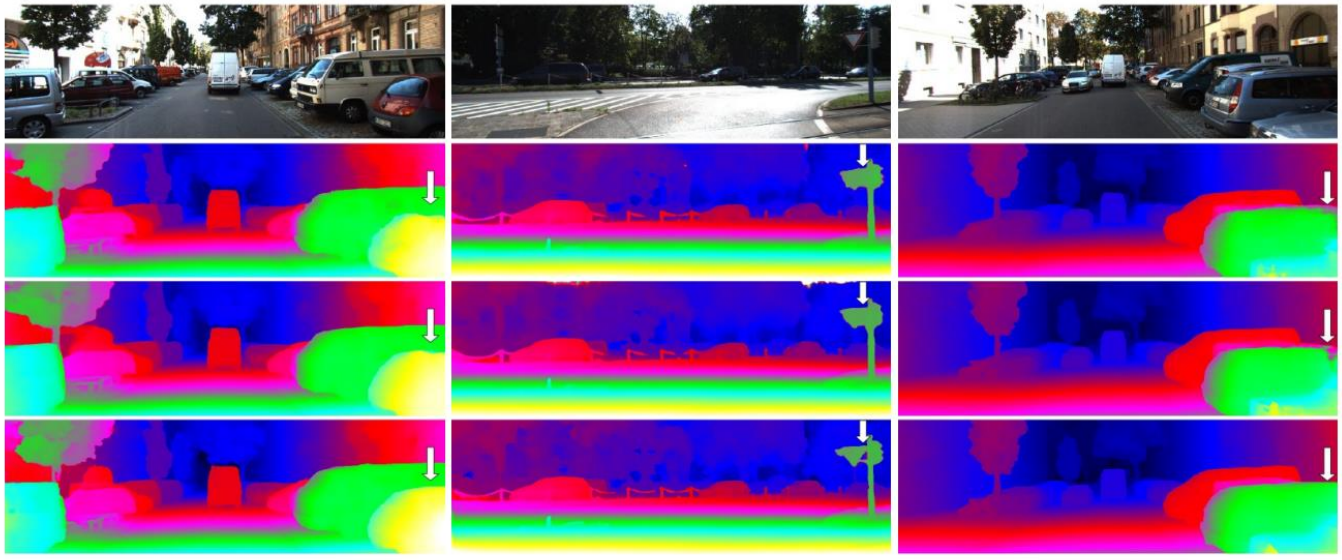


Figure 8: Qualitative results on the test set of KITTI2015 [32]. **Top row**: Input images, **Second row**: Results of PSMNet [3]. **Third row**: Results of GwcNet [15]. **Bottom row**: Results of GwcNet with cascade cost volume (GwcNet+Ours).

Paper_Cascaded MVS

- **Experiment Results_Stereo Matching**

	>1px	>2px.	>3px	EPE	Mem.
PSMNet [3]	9.46	5.19	3.80	0.887	6871
PSMNet+Ours	7.44	4.61	3.50	0.721	4124
GwcNet [15]	8.03	4.47	3.30	0.765	7277
GwcNet+Ours	7.46	4.16	3.04	0.649	4585
GANet11 [56]	-	-	-	0.95	6631
GANet11+Ours	11.0	5.97	4.28	0.90	5032

Table 4: Quantitative results of different stereo matching methods with and without cascade cost volume on Scene Flow dataset [30]. Accuracy, GPU memory consumption and run-time are included for comparisons.

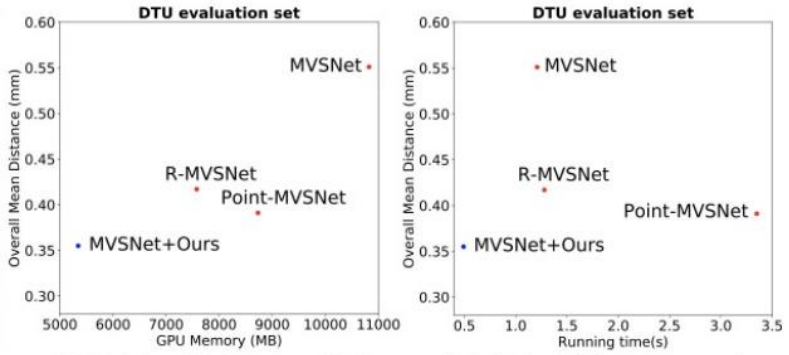
EPE: End point error
Disparity outlier percentage(%)

Methods	All (%)			Noc (%)		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all
DispNetC [30]	4.32	4.41	4.34	4.11	3.72	4.05
GC-Net [22]	2.21	6.16	2.87	2.02	5.58	2.61
CRL [34]	2.48	**3.59**	2.67	2.32	**3.12**	2.45
iResNet-i2e2 [27]	2.14	3.45	2.36	1.94	3.20	2.15
SegStereo [49]	1.88	4.07	2.25	1.76	3.70	2.08
PSMNet [3]	1.86	4.62	2.32	1.71	4.31	2.14
GwcNet [15]	1.74	3.93	2.11	1.61	3.49	1.92
GwcNet+Ours	**1.59**	4.03	**2.00**	**1.43**	3.55	**1.78**

Table 5: Comparison of different stereo matching methods on KITTI2015 benchmark [32].

D1: Percentage of stereo disparity outliers in first frame.

Conclusion



(e) Relationship between GPU memory and accuracy

(f) Relationship between running time and accuracy

- **Total Conclusion**

1. Save memory and running time
2. Decompose the single cost volume into a cascade formulation of multiple stages
3. Narrow the depth range of each stage and reduce the total number of hypothesis planes by utilizing the depth map from the previous stage
4. Use the cost volumes of higher spatial resolution to generate the outputs with finer details.

- **Limitations**
  - There is no clear difference between reconstruction picture compared to other techniques.
  - But you can see that they are absolutely superior in numerical figure.

Reference

1. Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In CVPR, 2018, pages 5410–5418, 2018.
2. Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In ICCV, 2019, 2019.
3. Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In CVPR, 2019, pages 3273–3282, 2019.
4. Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In ICCV, 2017, pages 66–75, 2017.
5. Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, ´ Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In CVPR, 2017, pages 2117– 2125, 2017.
6. Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In ECCV, 2018, pages 767–783, 2018.

Reference

7. Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multiview stereo depth inference. In CVPR, 2019, pages 5525– 5534, 2019.
8. Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In CVPR, 2015, pages 1592–1599, 2015.
9. Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-toend stereo matching. In CVPR, 2019, pages 185–194, 2019.