

# MULTI-TASK LEARNING OF DEPTH FROM TELE AND WIDE STEREO IMAGE PAIRS

---

Mostafa El-Khamy (presenter),  
Xianzhi Du, Haoyu Ren, Jungwon Lee  
SOC R&D, Samsung Semiconductor (SSI), San Diego, CA

*2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019*

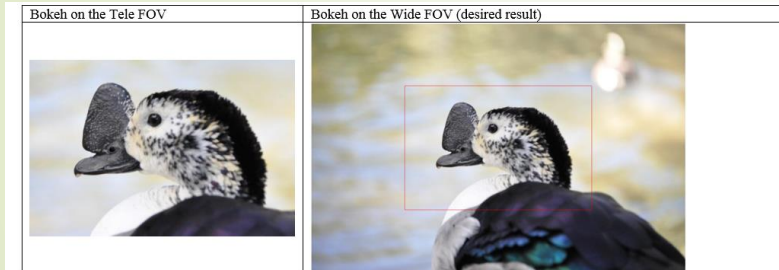
# Tele-Wide Stereo Matching: Motivation

- ▶ Recent multi-camera systems have **camera lenses that are chosen to have different focal lengths** to have good resolutions at different zoom ratios
  - ▶ Example: Dual Camera Phone:
    - ▶ First Camera has 26mm wide angle lens, and 2<sup>nd</sup> Camera is 52mm telephoto lens
- ▶ This work assumes the scenario when the scene is captured by two cameras with different field of views
  - ▶ The left lens has **1x zoom** and has a **wide-angle FOV** labeled as Wide FOV (WFOV, also called union FOV)
  - ▶ The right lens has **2x zoom** and has a narrower **telephoto FOV** (TFOV, also called overlapping FOV) which is centered in the WFOV

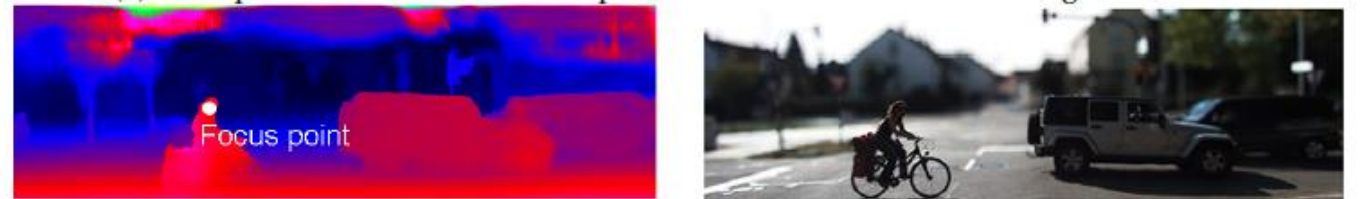


# Tele-Wide Stereo Matching: Definition

- ▶ We introduce the Tele-Wide Stereo Matching (TW-SM) problem:
- ▶ **TW-SM: Estimate the inverse depth (disparity) for the union WFOV, while leveraging the stereo information from the overlapping TFOV**
- ▶ Conventional stereo matching methods can only estimate the disparity for the Tele FOV
- ▶ With TW-SM, one can render the Bokeh for the full WFOV:



(a) Example of a tele-wide stereo input with a left Wide FOV and a right Tele FOV.

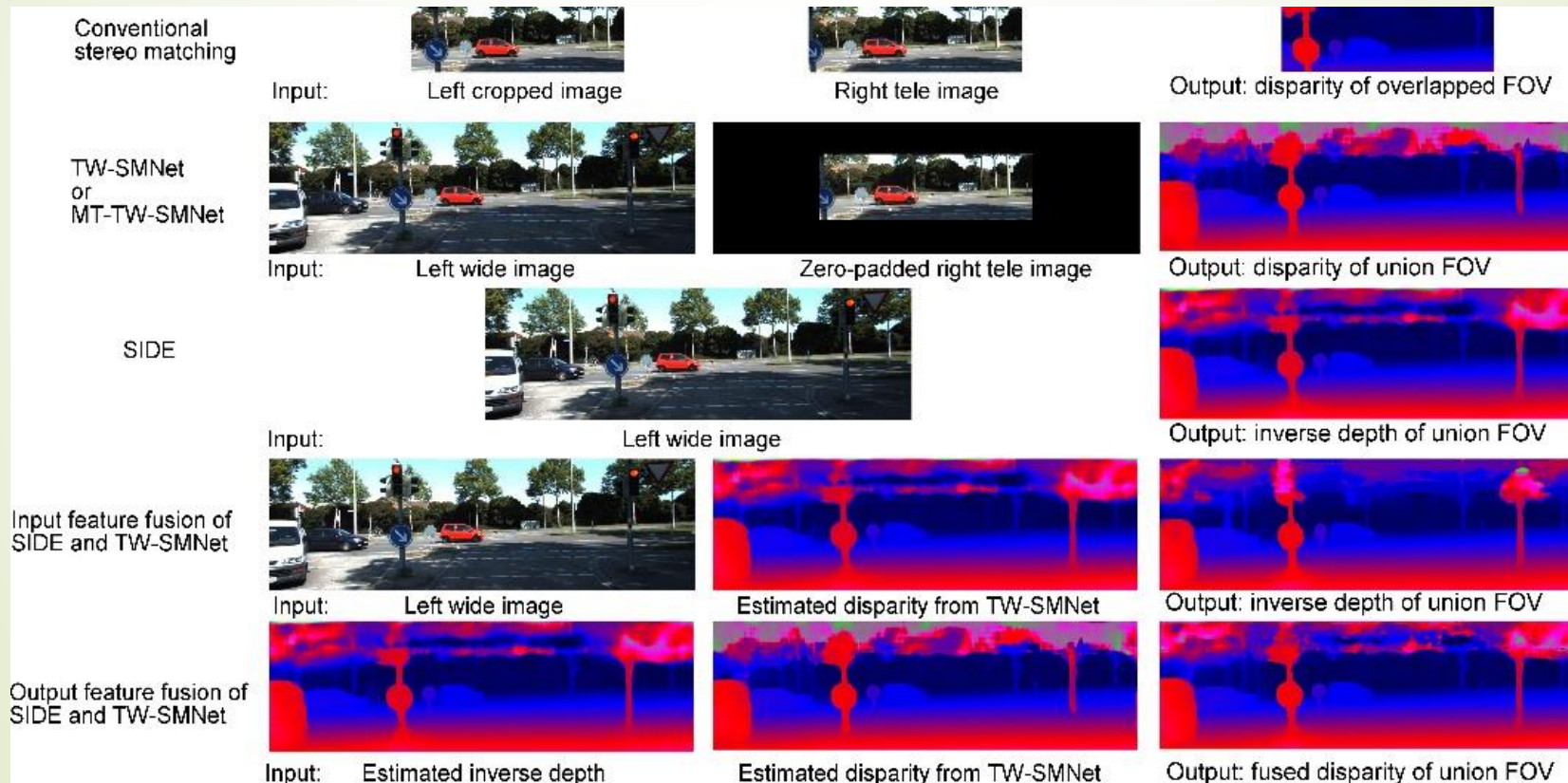


(b) The estimated tele-wide disparity map and the synthesized full-FOV Bokeh.

Tele-wide stereo matching on a KITTI example, to generate wide FOV Bokeh

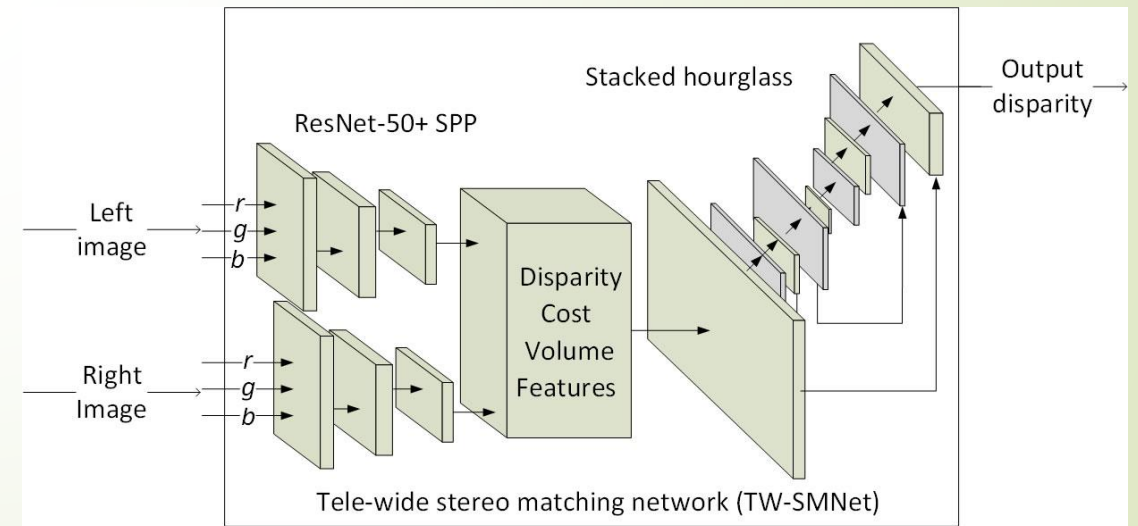
# TW-SM: Proposed Approaches

- ▶ We propose a unified architecture for deep neural networks that can perform depth estimation for the union of the field of views from 2 or more cameras, rather than for the overlapping intersection of FOVs only:
  - Single image inverse depth estimation (SIDE) network
  - Stereo matching network (TW-SMNet)
  - Multi-task network MT-TW-SMNet
  - Input/output feature fusion of above architectures (More recent than ICIP paper)



# TW-SMNet: TeleWide Stereo Matching Network

- ▶ Stereo Matched Disparity Estimation (SMDE) estimates the disparity for the union WFOV
- ▶ TW-SMNet(T) is Tele-Tele stereo matching which uses the cropped Tele regions only
  - ▶ TW-SMNet(T) is trained to output the disparity for the overlapping TFOV region only
- ▶ TW-SMNet(W) takes as input the left Wide image and the right Tele image after appropriate scaling, rectification, and zero padding
  - ▶ TW-SMNet(W) is trained to output the disparity for the union WFOV region
    - ▶ TW-SMNet follows the spatial pyramid pooling (SPP) architecture of PSMNet
    - ▶ SPP aggregates the global context from a ResNet50 feature extractor at multiple scales and locations
    - ▶ The cost volume is constructed by concatenating left and right features at different disparity shifts, till max disparity  $D$
    - ▶ A 4D **stacked hourglass (SHG)** CNN aggregates information from the cost volume ( $D+1, W, H, Ch$ )
    - ▶ The disparity is optimized by minimizing the Huber loss of the error between the disparity estimated using soft regression and the ground truth disparity

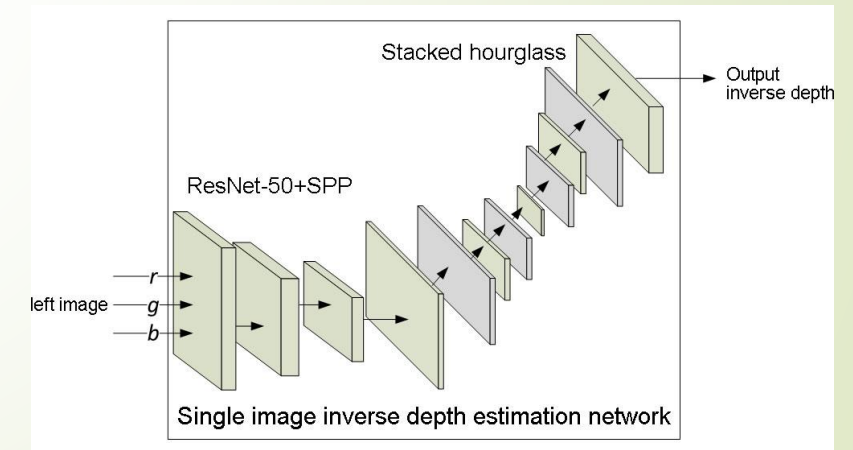
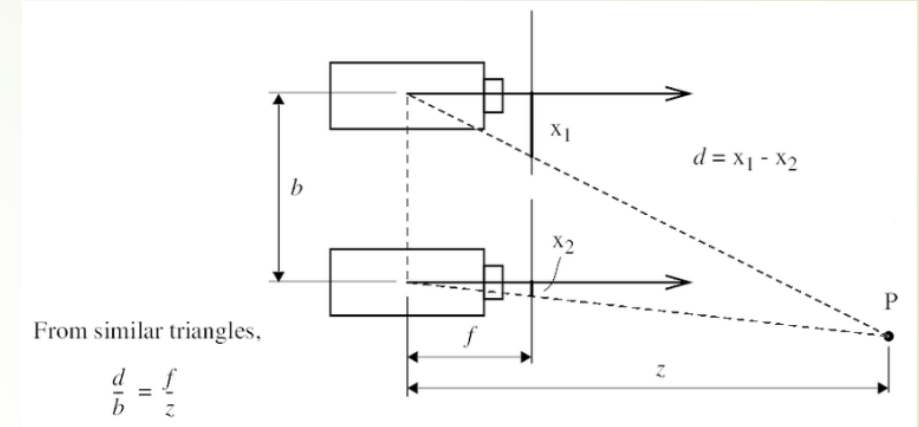


# SHG-SIDE: Wide FOV Single-Image Inverse Depth Estimation (SIDE)

- ▶ TW-SMNet's accuracy in the surrounding region is worse than that in the overlapped TFOV, due to the missing stereo information
- ▶ With the knowledge of the camera baseline  $b$  and focal length  $f$ , the disparity  $d$  is proportional to the inverse depth  $1/z$  of the subject by

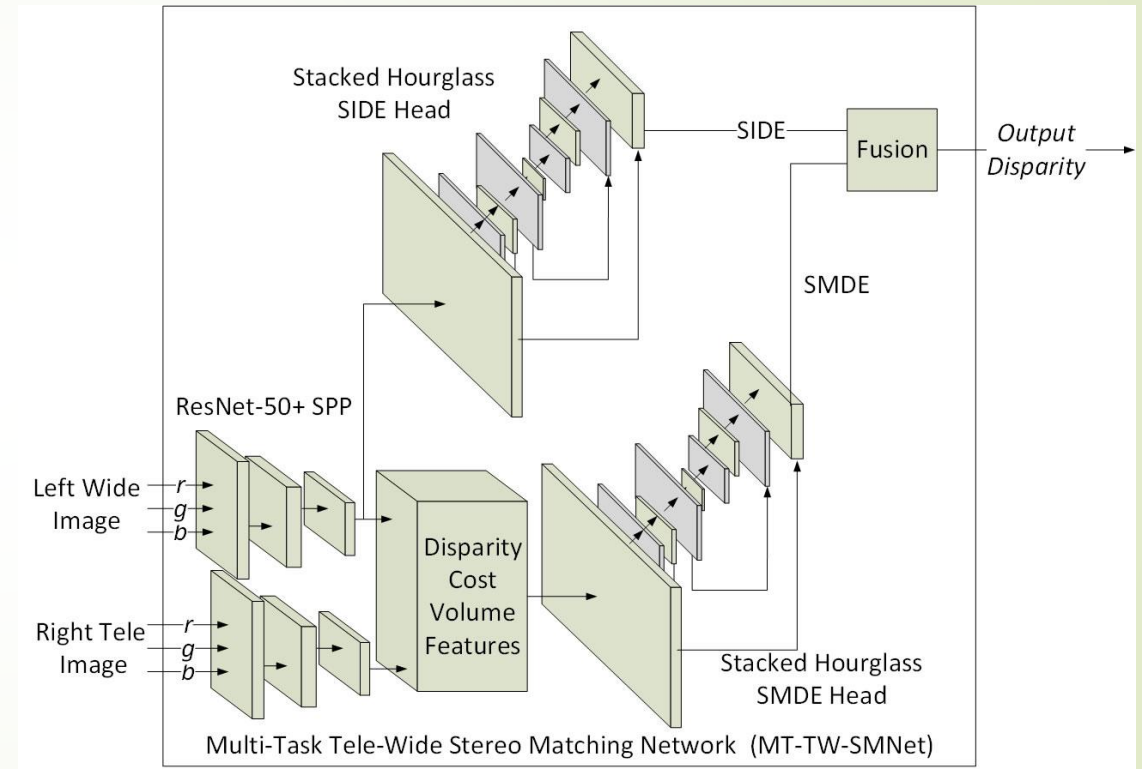
$$\frac{1}{z} = \frac{d}{fb}$$

- ▶ SHG-SIDENet: Using the left WFOV image only, we modify the TW-SMNet to perform Single-Image inverse depth estimation
  - ▶ The cost volume is removed, and the 4D SHG is replaced by a 3D SHG
  - ▶ The network is trained using soft regression to the inverse depth of the full WFOV



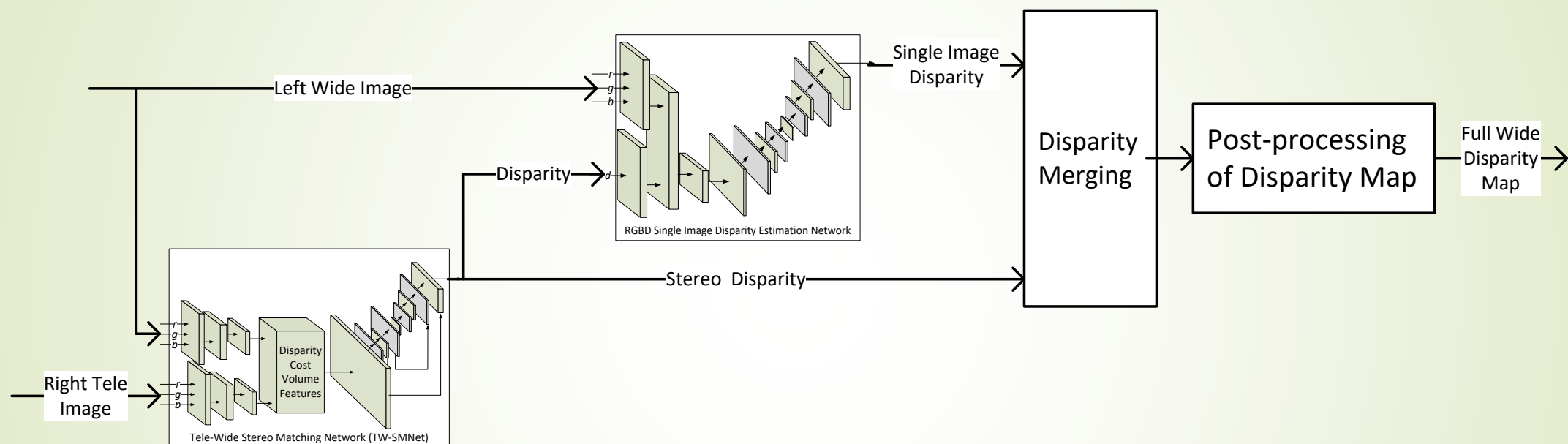
# MT-TW-SMNet: Multitask Tele-Wide Stereo Matching Network

- ▶ TW-SMNet is more accurate in the Tele FOV
- ▶ SHG-SIDENet is more accurate in the surrounding part (around the TFOV) of the Wide FOV
- ▶ The MT-TW-SMNet is trained end-to-end to optimize the objectives of both the SHG-SIDE and the TW-SM
  - ▶ The ResNet50 + SPP feature extractor is shared between both objectives
  - ▶ The loss function is a weighted combination of both loss functions to minimize the errors from the classification-based regressions



# TW-SM: Fusion (More Recent Work)

- General input-output fusion architecture

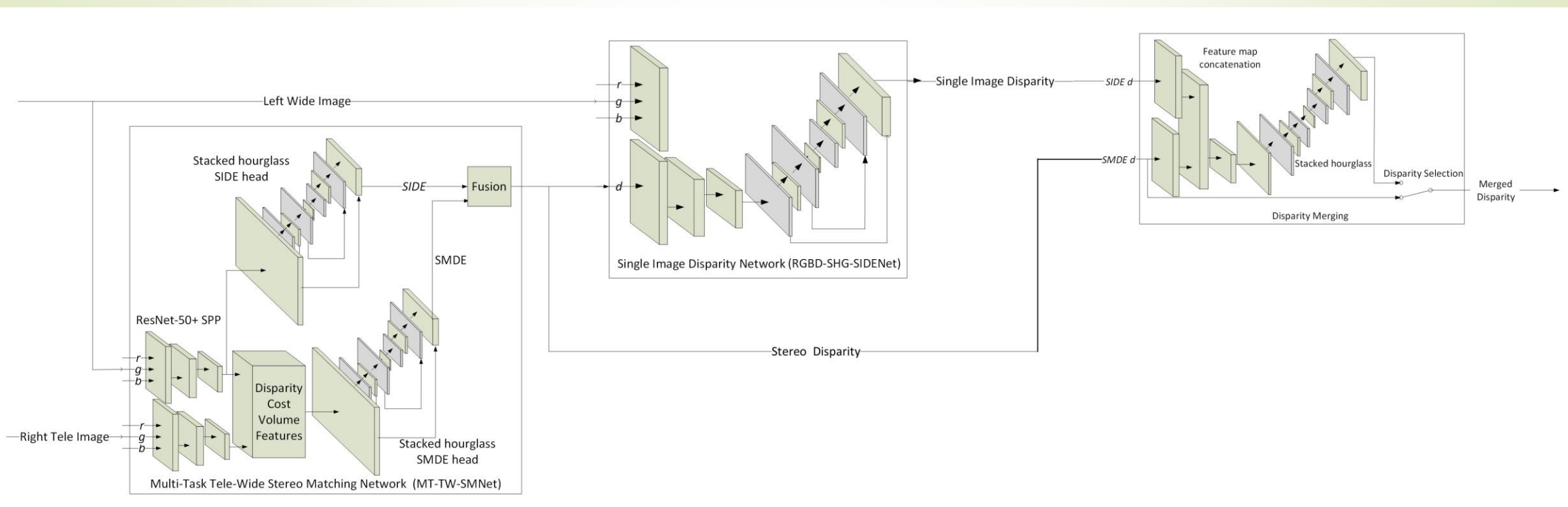


- Input Feature Fusion
- The output disparity of TW-SMNET (TW-SMNet(T) or TW-SMNet(W) or MT-TW-SMNet) is concatenated as an additional input feature to the SIDE, denoted as SIDE-RGBD
  - The additional input feature guides the RGBD-SIDE to obtain more accurate estimates in the overlapping TFOV region



# TW-SM: Fusion (More Recent Work)

- ▶ Putting it all together with input and output (decision) fusion
- ▶ Deep Network Fusion of the disparity maps: The output fusion network is trained with the classification based robust regression to the ground truth disparity of the WFOV, and uses a SHG architecture
  - ▶ Disparity selection: e.g. Select the MT-TW-SMNet output in the Tele region and SIDE output in the surrounding region



# Results

- ▶ The TW-SMDE networks are trained on the SceneFlow dataset and the KITTI dataset
- ▶ The datasets were processed to generate the Tele and Wide Stereo image pairs
  - ▶ The center cropped region of the right image, of half the original width and height, is used as the Right Tele Image. The left image is the left wide Image.

TABLE 1: Error rate of tele-wide disparity estimation networks on the KITTI stereo 2015 validation dataset; 'cen' stands for the 'center' TFOV region, and 'sur' stands for the region in the WFOV 'surrounding' the TFOV.

Model name	type	left image	right image	error-all(%)	error-cen(%)	error-sur (%)
DORN	single image	wide	N/A	17.55	11.96	20.68
SHG-SIDENet	single image	wide	N/A	12.62	7.31	15.80
TW-SMNet(T)	stereo	tele	tele	N/A	1.68	N/A
TW-SMNet(W)	stereo	wide	tele	13.10	1.86	19.63
MT-TW-SMNet	stereo	wide	tele	12.70	1.94	18.99




TABLE 2: End-point error of tele-wide disparity estimation networks on the SceneFlow test set.

Model name	type	left image	right image	error-all(pixels)	error-cen(pixels)	error-sur (pixels)
SHG-SIDENet	single image	wide	N/A	7.47	8.32	7.20
TW-SMNet(T)	stereo	tele	tele	N/A	1.28	N/A
TW-SMNet(W)	stereo	wide	tele	5.79	1.88	7.10
MT-TW-SMNet	stereo	wide	tele	5.61	1.62	7.08

# Results

## ► The results on KITTI Stereo 2015 Test Set

- Tele-Wide SM ranks fairly compared to other methods that do Wide-Wide stereo matching on the WFOV stereo image pair
- Compared to SIDE, MT-TWSMNet improved the overall error rate from 20.7% to 15.6%
- Fusion improved the overall error rate from 15.6% to 11.96%

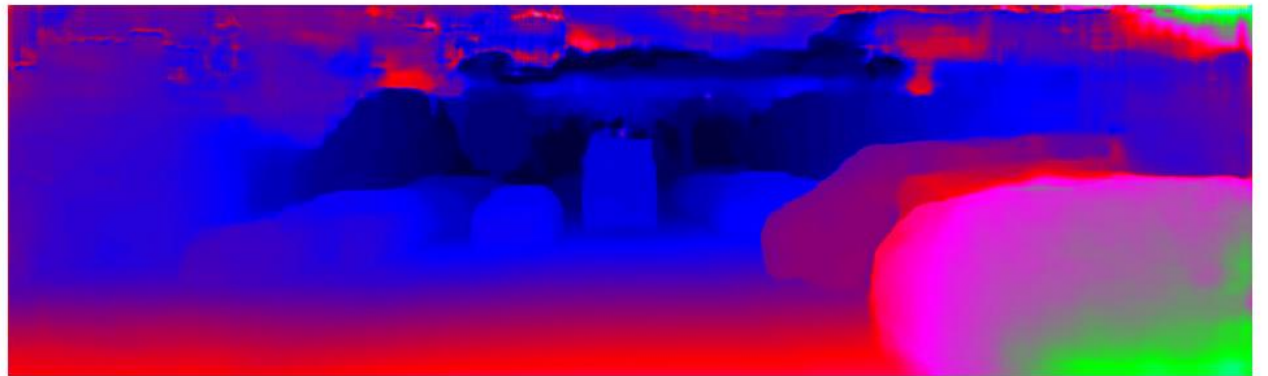
H. Hirschmüller: <a href="#">Stereo processing by semiglobal matching and mutual information</a> . PAMI 2008.						
177	<a href="#">TW-SMNet</a>			11.92 %	12.16 %	11.96 % 100.00 %
178	<a href="#">SDM</a>			9.41 %	24.75 %	11.96 % 62.56 %
J. Kostkova: <a href="#">Stratified dense matching for stereopsis in complex scenes</a> . BMVC 2003.						
179	<a href="#">SGM&amp;FlowFie+</a>			11.93 %	20.57 %	13.37 % 81.24 %
R. Schuster, C. Bailer, O. Wasenmüller and D. Stricker: <a href="#">Combining Stereo Disparity and Optical Flow for Basic Stereo Matching</a>						
180	<a href="#">GCSE</a>		<a href="#">code</a>	11.64 %	27.11 %	14.21 % 100.00 %
J. Cech, J. Sanchez-Riera and R. Horaud: <a href="#">Scene Flow Estimation by growing Correspondence Seeds</a> . CVPR 2011.						
181	<a href="#">MT-TW-SMNet</a>			15.47 %	16.25 %	15.60 % 100.00 %
182	<a href="#">Mono-SF</a>			14.21 %	26.94 %	16.32 % 100.00 %
183	<a href="#">CostFilter</a>		<a href="#">code</a>	17.53 %	22.88 %	18.42 % 100.00 %
C. Rhemann, A. Hosni, M. Bleyer, C. Rother and M. Gelautz: <a href="#">Fast Cost-Volume Filtering for Visual Correspondence</a>						

Percentage of erroneous pixels

Model Name	Background	Foreground	All Pixels
SIDENet	20.19	23.44	20.73
MT-TW-SMNet	15.47	16.25	15.60
MT-TW-SM Fusion	11.92	12.16	11.96

# Results

- ▶ Example disparity map output using MT-TW-SM on KITTI Stereo 2015 Test set
- ▶ Two stereo inputs,
  - ▶ Left Wide image
  - ▶ Right Tele Image
- ▶ The generated disparity map is for the WFOV



# Application

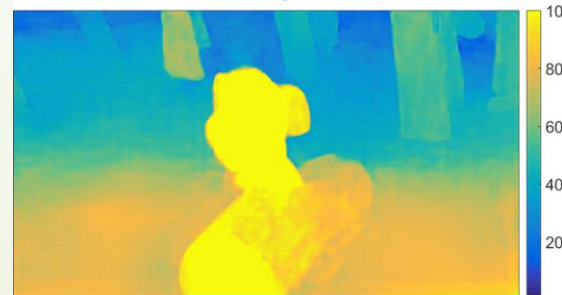
- ▶ Synthesizing the Bokeh effect on the full Wide FOV
  - ▶ We show an example of TW-SM Disparity Estimation (TW-SMDE) trained on SceneFlow
  - ▶ Using a Tele and Wide stereo image pair, Bokeh is rendered on the full WFOV



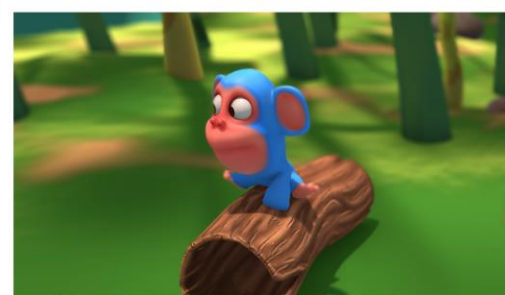
Left wide image



Right tele image



Our estimated disparity



Our Bokeh result

# Conclusions

- ▶ We introduced the problem of Tele-Wide depth estimation:
  - ▶ Estimate the inverse depth (disparity) for the union WFOV, while leveraging the stereo information from the overlapping TFOV
- ▶ We proposed a general framework and deep learning solutions to solve this problem:
  - ▶ Estimate the inverse depth for the WFOV from the Wide Image (20.7% error rate on KITTI)
  - ▶ Estimate depth or disparity for the WFOV (or the TFOV only) from both the Wide Image and the Tele image
- ▶ We proposed a Multi-task Deep Neural Network that attempts to solve these 2 problems together (15.6% error rate on KITTI)
- ▶ We proposed Input and Output fusion networks to further refine the results and leverage the different accuracy gains of Tele-Wide stereo matching and inverse depth estimations (11.9% error rate on KITTI)
- ▶ Using these one can generate full FOV Bokeh from a Tele and Wide stereo image pair



(a) Example of a tele-wide stereo input with a left Wide FOV and a right Tele FOV.



(b) The estimated tele-wide disparity map and the synthesized full-FOV Bokeh.

Thank You!