Monocular-Depth-Estimation:

The influence of image augmentation on state-of-the-art depth estimators

Leonardo Pohl

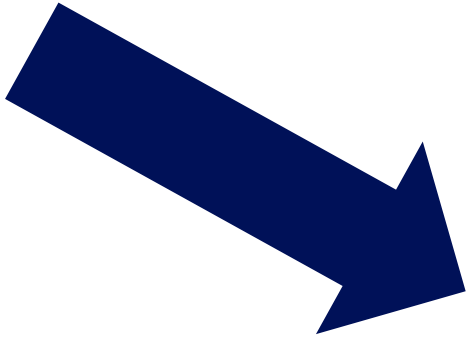
Universiteit
Leiden
The Netherlands

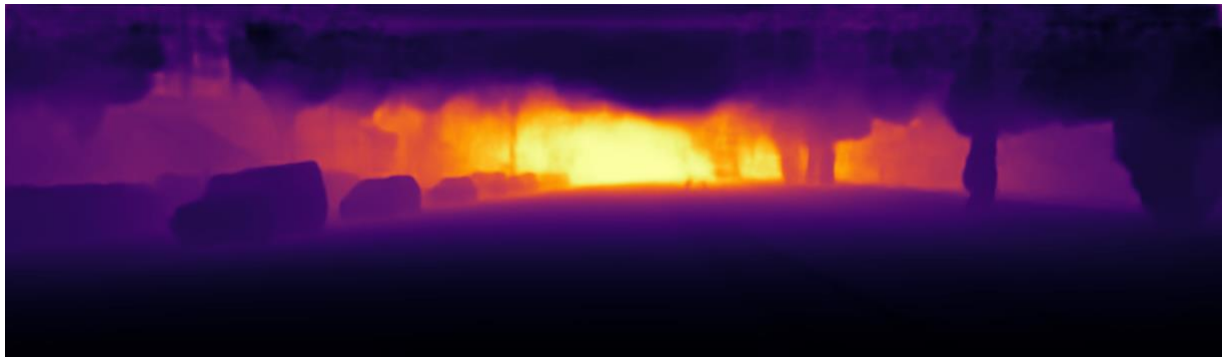
Humans

- Evolution has improved certain human features to perfection
- Is this animal close enough to pose a threat?
- Can I reach this fruit?
- Humans had millions of years
- Modern Computers have existed for less than 100 years
- Humans are impatient

Monocular-Depth-Estimation

- Extract the depth information from a single RGB image
- Highly researched topic in Computer Vision
- Recently a lot of papers have been released





Estimators

- Most estimators use machine learning or neural networks

1. GLPN: Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth
2. NeWCRFs: Neural Window Fully-connected CRFs for Monocular Depth Estimation

GLPN: Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth [1]

- GLPN extracts global and local features to estimate the depth of the image
- Depth specific image augmentation Vertical CutDepth is applied [2]
  - Part of the depth image is passed in cut into the image during training

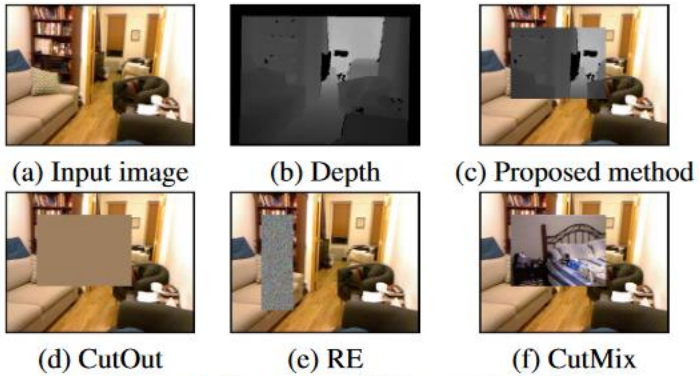


Figure 1. Examples of data augmentation

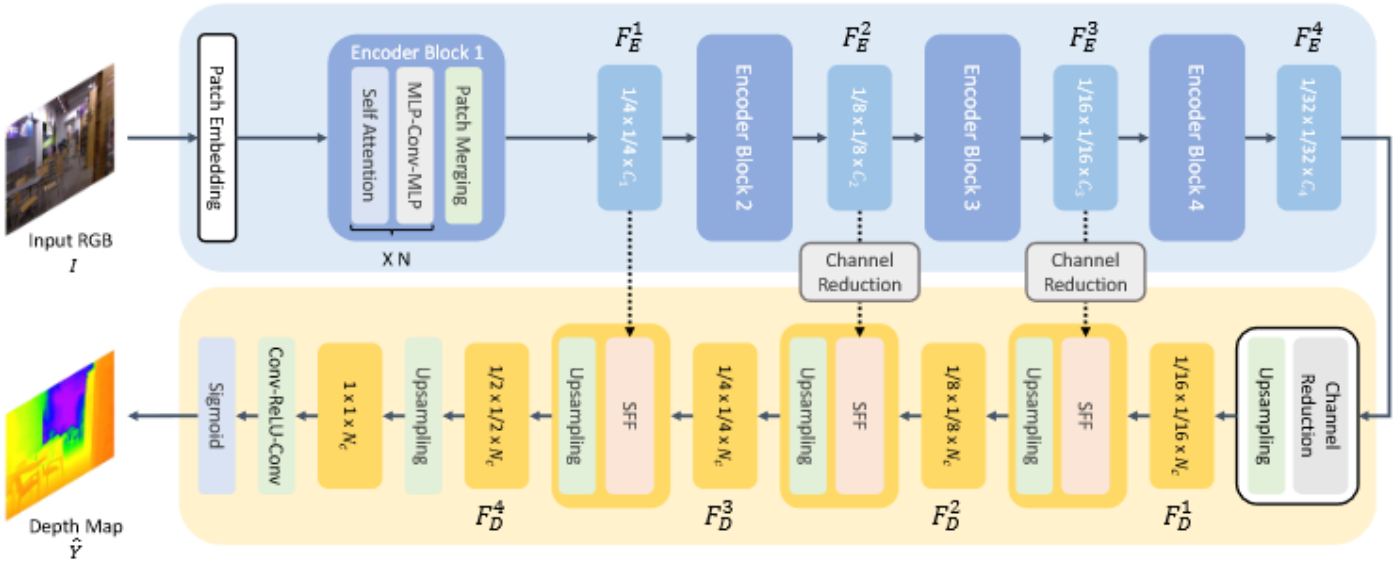


Figure 1: Overall architecture of the proposed network. The main components of the architecture are the encoder, decoder, and skip connections with feature fusion modules.

[1] - Kim, D., Ga, W., Ahn, P., Joo, D., Chun, S., & Kim, J. (2022). Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth. *arXiv preprint arXiv:2201.07436*.
[2] - Ishii, Y., & Yamashita, T. (2021). CutDepth:Edge-aware Data Augmentation in Depth Estimation (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2107.07684

NeWCRFs: Neural Window Fully-connected CRFs for Monocular Depth Estimation[1]

- Use Conditional Random Field to regress the depth map [2]
- Split the input into windows and perform FC-CRFs optimisation on it [3]

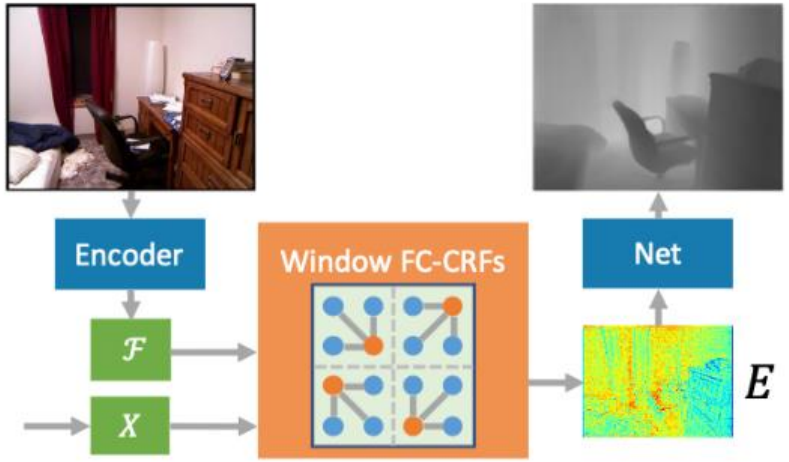


Figure 1. The neural window fully-connected CRFs take image feature $\mathcal{F}$ and upper-level prediction X as input, and compute the fully-connected energy E in each window, which is then fed to the networks to output an optimized depth map.

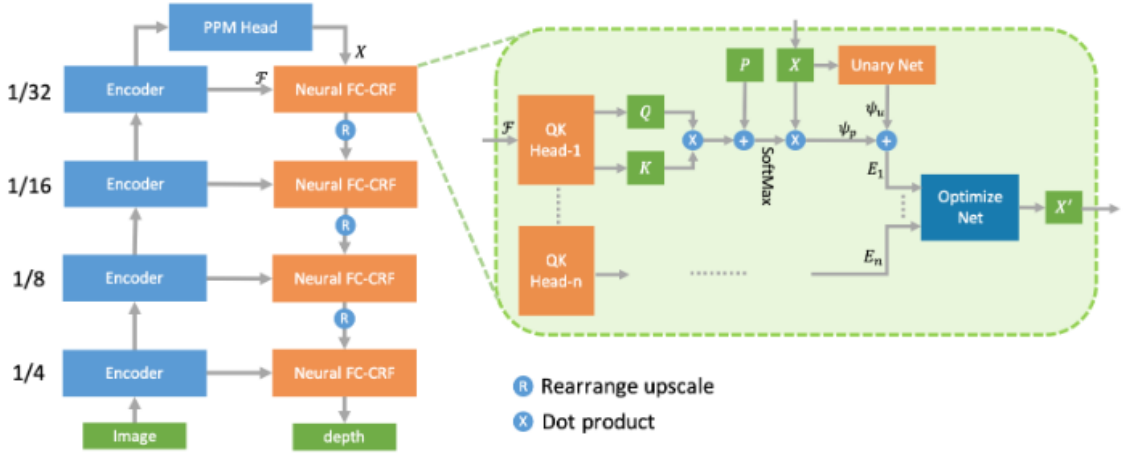


Figure 3. Network structure of the proposed framework. The encoder first extracts the features in four levels. A PPM head aggregates the global and local information and makes the initial prediction X from the top image feature $\mathcal{F}$. Then in each level, the neural window fully-connected CRFs module builds multi-head energy from X and $\mathcal{F}$, and optimizes it to a better prediction X'. Between each level a rearrange upscale is performed considering the sharpness and network weight.

[1] - Yuan, W., Gu, X., Dai, Z., Zhu, S., & Tan, P. (2022). NeW CRFs: Neural Window Fully-connected CRFs for Monocular Depth Estimation (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2203.01502
[2] - Zheng, Shuai & Jayasumana, Sadeep & Romera-Paredes, Bernardino & Vineet, Vibhav & Su, Zhizhong & Du, Dalong & Huang, Chang & Torr, Philip. (2015). Conditional Random Fields as Recurrent Neural Networks.
[3] - Zhang, Bin. (2018). Fully Connected Conditional Random Fields for High-Resolution Remote Sensing Land Use/Land Cover Classification with Convolutional Neural Networks. Remote Sensing. 10. 10.3390/rs10121889.

Datasets

- Somehow this has to be tested:
- Datasets have to include a depth ground truth
- Multiple Datasets exist
- I picked the two most used datasets
  - KITTI Eigen Split Dataset
  - NYU-Depth V2 Dataset

KITTI Dataset

- Collaboration between KIT and TTI (Karlsruhe and Toyota)
- Data accumulated by cameras mounted on a car
- The ground truth is calculated using a 360° Velodyne Laserscanner

[1] - Andreas Geiger, Philip Lenz, & Raquel Urtasun (2012). Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

NYU-Depth V2 Dataset

- Developed by NYU – New York University
- Data accumulated composed of frames of videos taken using a Kinect
- Various Indoor settings

[1] - Nathan Silberman, P., & Rob Fergus (2012). Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*.

Evaluation Metrics

- Root mean squared error
  - $RMSE = \sqrt{mean((gt - pred)^2)}$
- Relative squared and absolute error
  - $sq\_rel = mean\left(\frac{(gt-pred)^2}{gt}\right)$
  - $abs\_rel = mean\left(\frac{abs(gt-pred)}{gt}\right)$
- δ_1, δ_2 and δ_3
  - With $threshold = \max\left(\frac{pred}{gt}, \frac{gt}{pred}\right)$
  - $\delta_1 = mean(threshold < 1.25)$
  - $\delta_2 = mean(threshold < 1.25^2)$
  - $\delta_3 = mean(threshold < 1.25^3)$
- Scale-Invariant Error
  - $silog = \sqrt{mean\big((\log(gt) - \log(pred))^2\big) - mean(\log(gt) - \log(pred))^2}$

GLPN vs. NeWCFs

	silog	log10	rmse	rmse_log	abs_rel	sq_rel	d1	d2	d3
level_0									
GLPN	14.0430	0.0529	5.3478	0.1763	0.1384	1.1728	0.8745	0.9735	0.9912
NeWCFs	5.6374	0.0178	1.9663	0.0629	0.0405	0.1192	0.9860	0.9980	0.9996

(a) Experiment results of GLPN vs. NeWCFs on the KITTI Eigen Dataset.

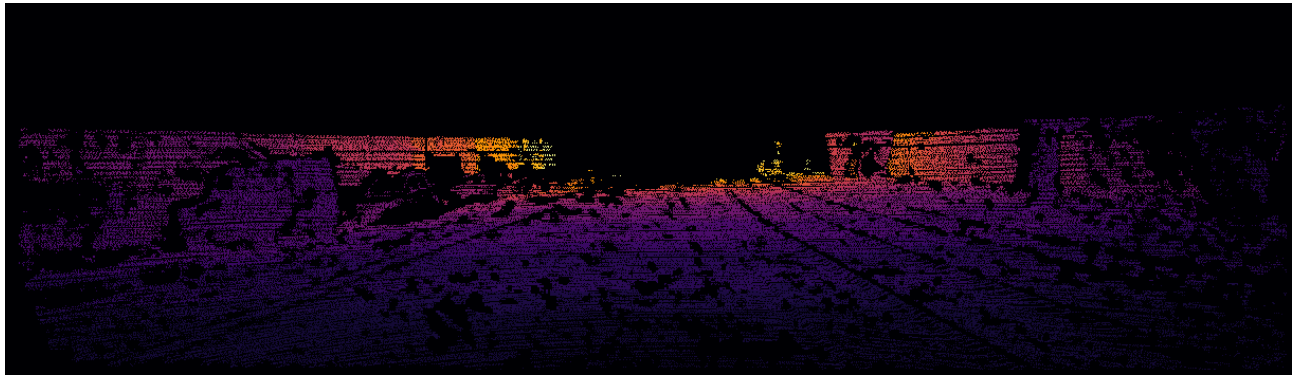
	silog	log10	rmse	rmse_log	abs_rel	sq_rel	d1	d2	d3
level_0									
GLPN	3.0834	0.0117	0.0885	0.0369	0.0273	0.0035	0.9982	0.9997	0.9999
NeWCFs	3.8978	0.0150	0.1722	0.0474	0.0343	0.0078	0.9961	0.9994	0.9998

(b) Experiment results of GLPN vs. NeWCFs on the NYU-Depth V2 Dataset.

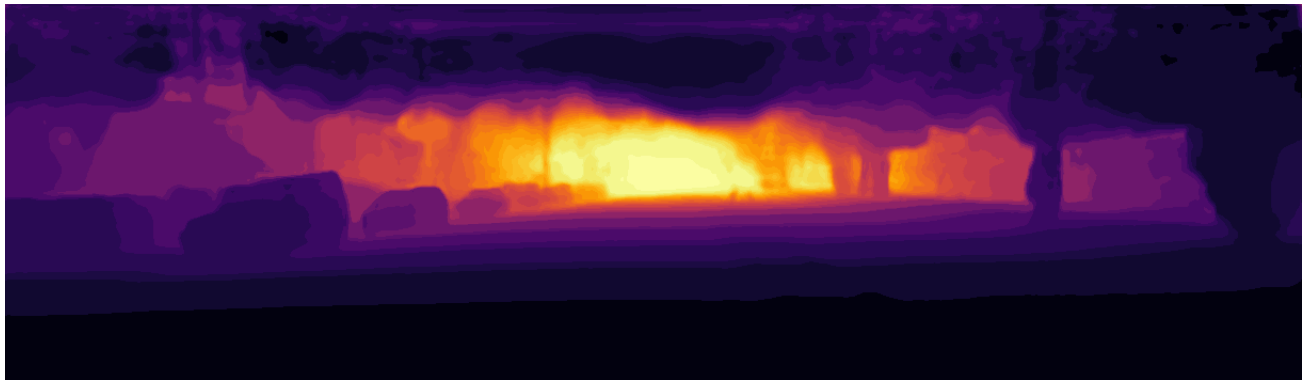
Figure 5. The results of the experiments without image augmentation

GLPN vs. NeWCFs

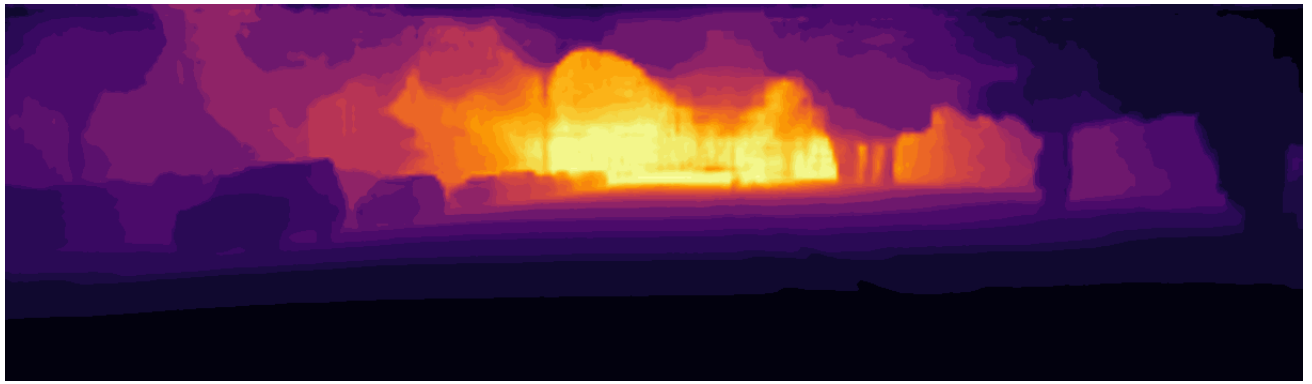
GT



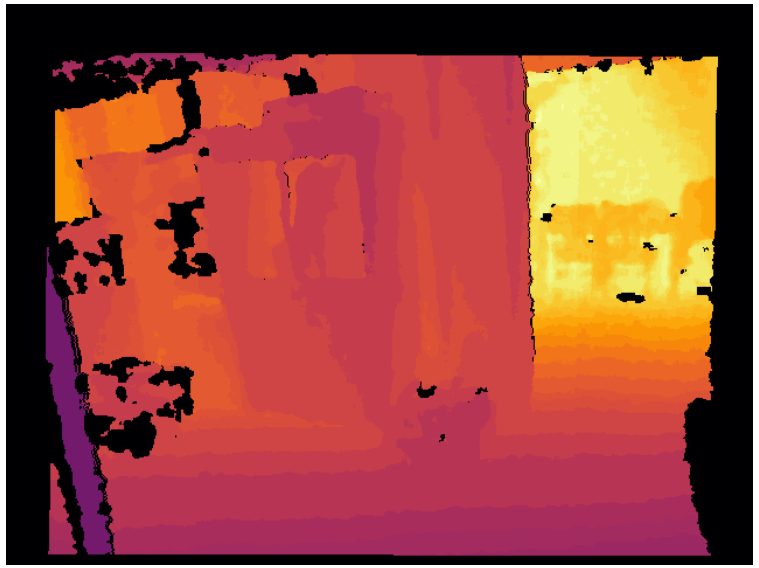
GLPN



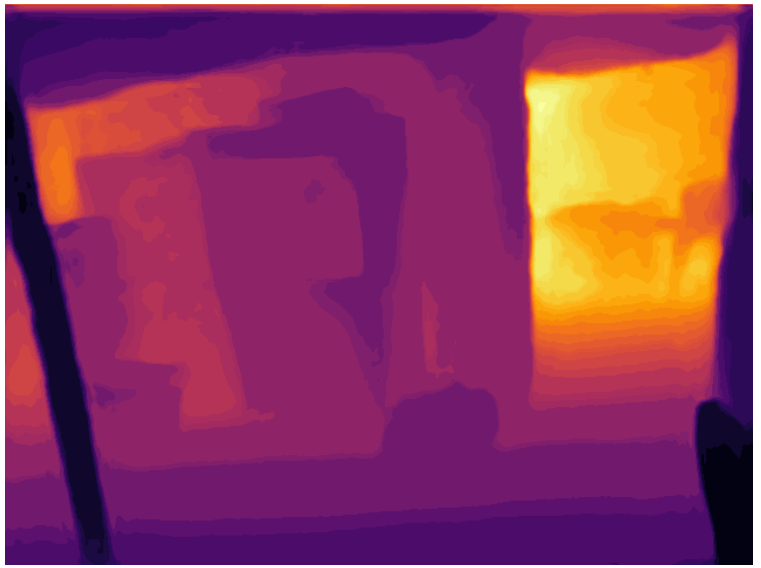
NeWCFs



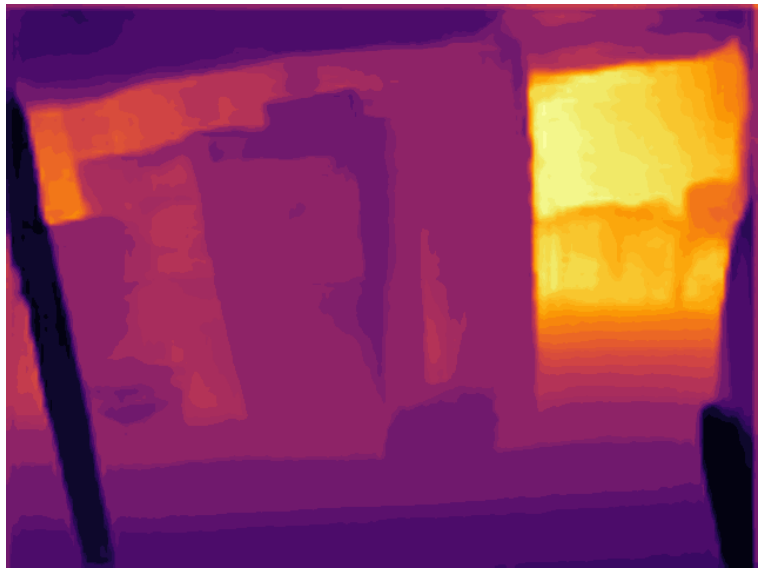
GLPN vs. NeWCFs



GT



GLPN



NeWCFs

Image augmentation

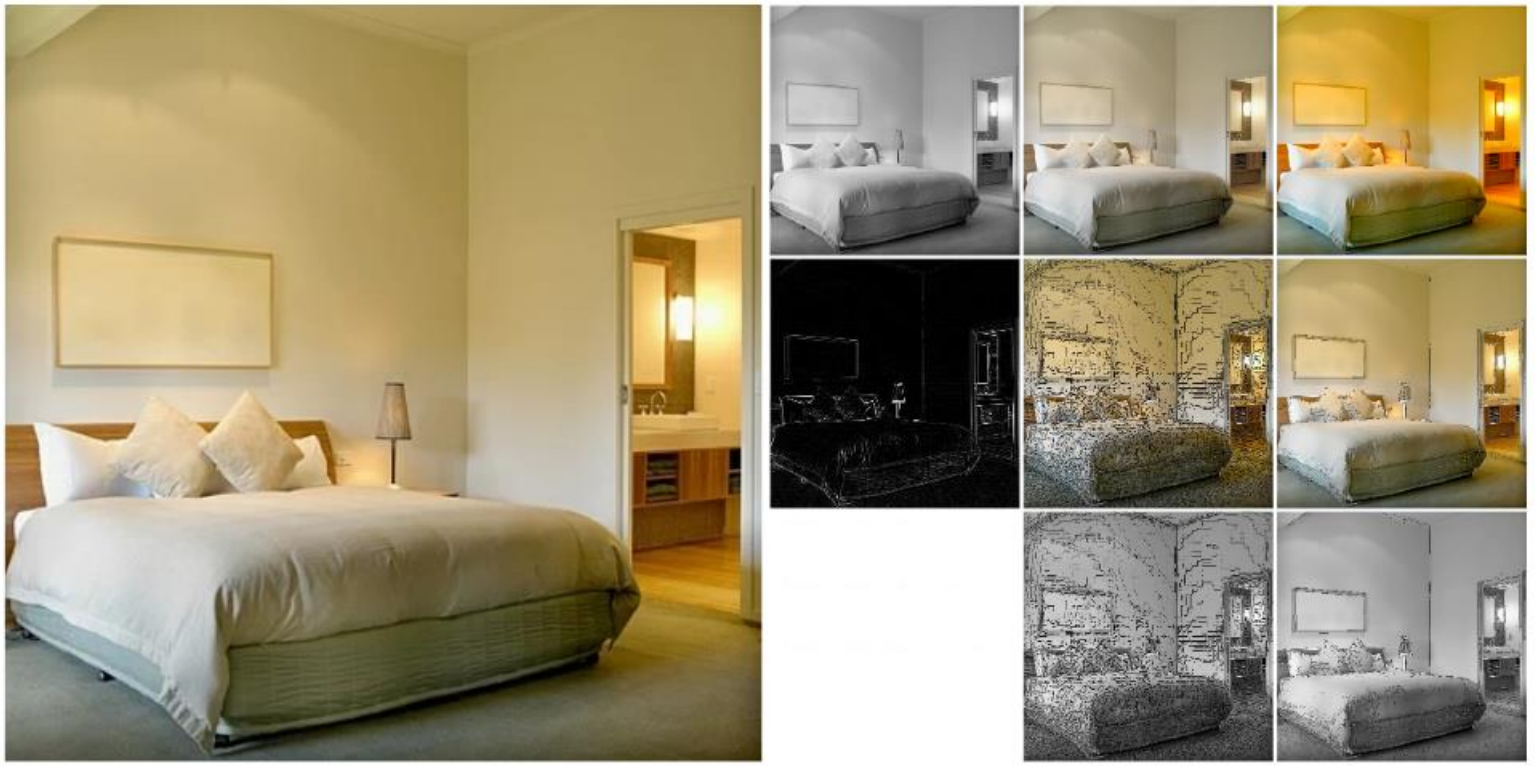


Figure 4. Different Data augmentation methods applied on the sample image which can be seen on the left. The images were augmented from left to write int the following ways: (a) Greyscaled, (b) Desaturated, (c) Oversaturated, (d) Edge Detection, (e) Colour Image overlayed with edges with > 2% white value, (f) Colour Image overlayed with edges with > 20% white value, (g) Greyscaled Image overlayed with edges with > 2% white value and (h) Greyscaled Image overlayed with edges with > 20% white value

Image Augmentation

GLPN with Image augmentation

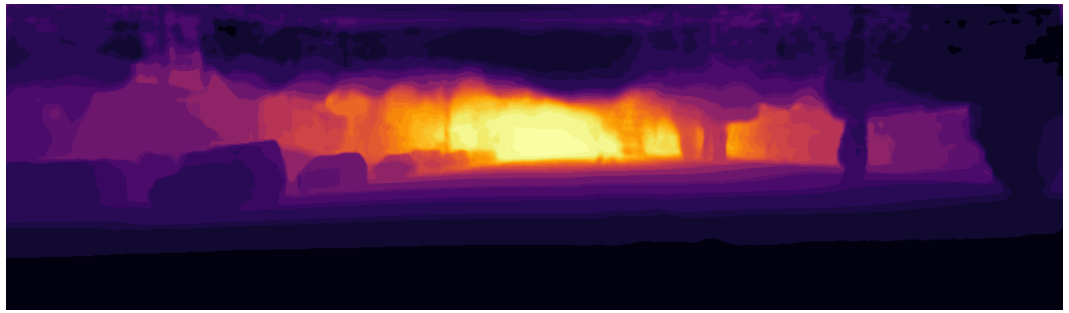
Preprocessing Type	silog	log10	rmse	rmse_log	abs_rel	sq_rel	d1	d2	d3
Original	14.0430	0.0529	5.3478	0.1763	0.1384	1.1728	0.8745	0.9735	0.9912
Greyscale	14.4444	0.0496	5.0262	0.1681	0.1259	0.9911	0.8831	0.9753	0.9923
Desaturated	14.0790	0.0526	5.3319	0.1757	0.1374	1.1667	0.8753	0.9736	0.9913
Oversaturated	13.9611	0.0528	5.2981	0.1753	0.1377	1.1473	0.8746	0.9740	0.9915
Edge Detection	16.9108	0.0501	5.2132	0.1802	0.1169	0.9004	0.8662	0.9618	0.9874
Edge Detection and Colour Threshold 5	15.3736	0.0469	4.9053	0.1653	0.1157	0.8883	0.8873	0.9747	0.9918
Edge Detection and Colour Threshold 50	14.5960	0.0510	5.1737	0.1723	0.1311	1.0804	0.8756	0.9744	0.9916
Edge Detection and Greyscale Threshold 5	17.0663	0.0508	5.2565	0.1812	0.1177	0.9184	0.8615	0.9624	0.9879
Edge Detection and Greyscale Threshold 50	15.2886	0.0499	5.1046	0.1709	0.1246	0.9919	0.8750	0.9735	0.9918

(a) Results of the experiments run on the KITTI Eigen Dataset

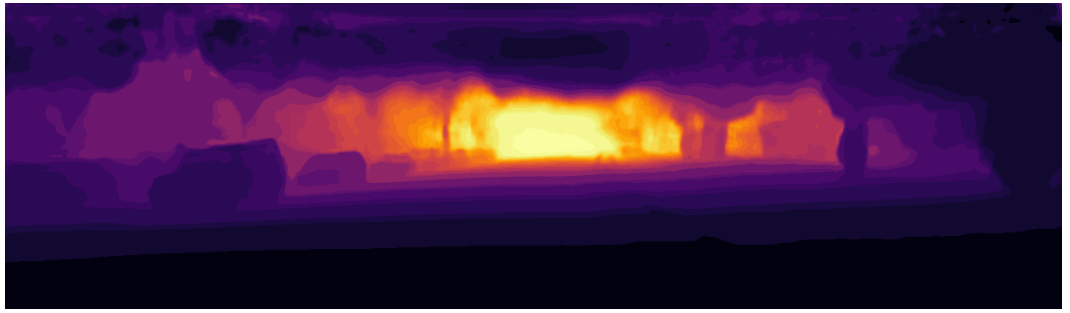
Preprocessing Type	silog	log10	rmse	rmse_log	abs_rel	sq_rel	d1	d2	d3
Original	3.0834	0.0117	0.0885	0.0369	0.0273	0.0035	0.9982	0.9997	0.9999
Greyscale	3.6161	0.0137	0.1009	0.0427	0.0320	0.0046	0.9975	0.9996	0.9999
Desaturated	3.1056	0.0118	0.0893	0.0372	0.0276	0.0035	0.9982	0.9997	0.9999
Oversaturated	3.1796	0.0120	0.0899	0.0378	0.0280	0.0036	0.9981	0.9997	0.9999
Edge Detection	17.5686	0.0825	0.5362	0.2336	0.1780	0.1311	0.6893	0.9072	0.9789
Edge Detection and Colour Threshold 5	14.8221	0.0619	0.3920	0.1819	0.1533	0.0834	0.8002	0.9606	0.9901
Edge Detection and Colour Threshold 50	4.4382	0.0189	0.1515	0.0581	0.0451	0.0114	0.9925	0.9986	0.9998
Edge Detection and Greyscale Threshold 5	19.1446	0.0819	0.5253	0.2359	0.2002	0.1422	0.6814	0.9202	0.9791
Edge Detection and Greyscale Threshold 50	5.8929	0.0228	0.1851	0.0720	0.0547	0.0181	0.9828	0.9968	0.9995

(b) Results of the experiments run on the NYU-Depth V2 Dataset

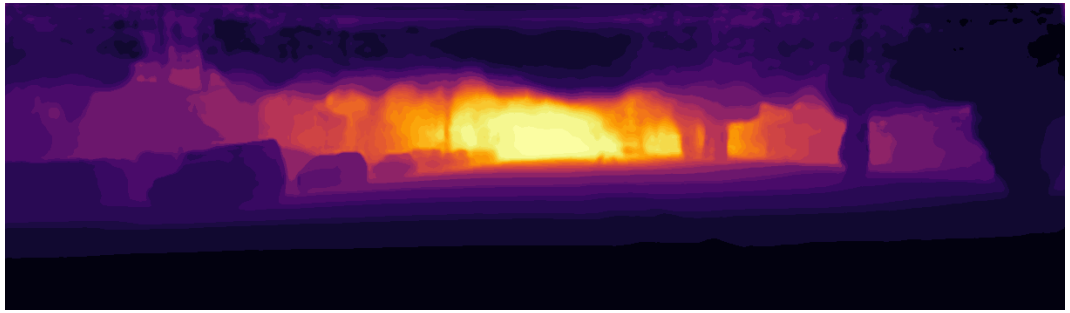
GLPN With Image Augmentation KITTI



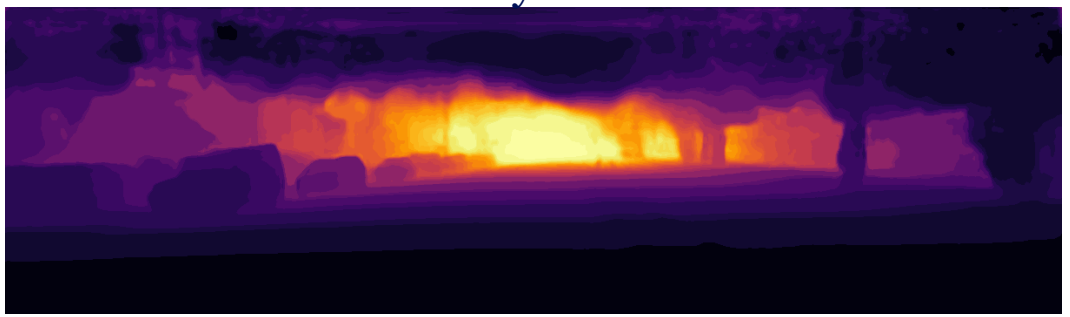
Original



Greyscale

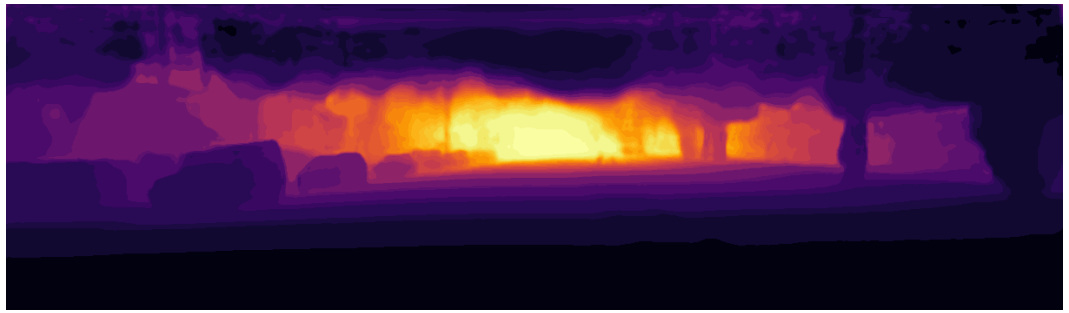


Oversaturated



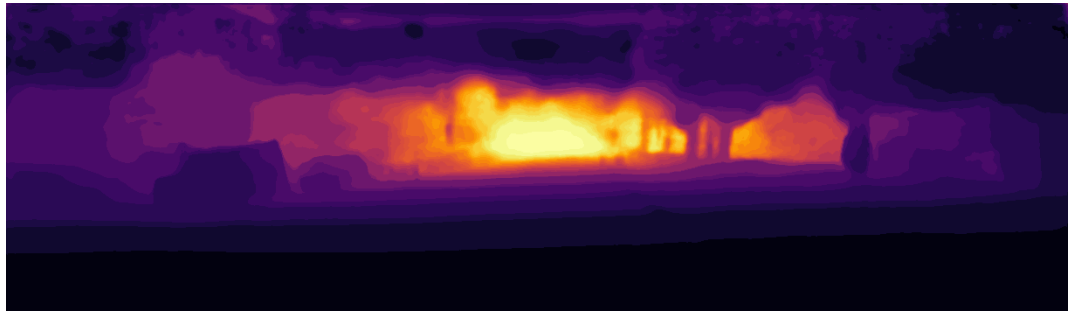
Desaturated

GLPN With Image Augmentation KITTI

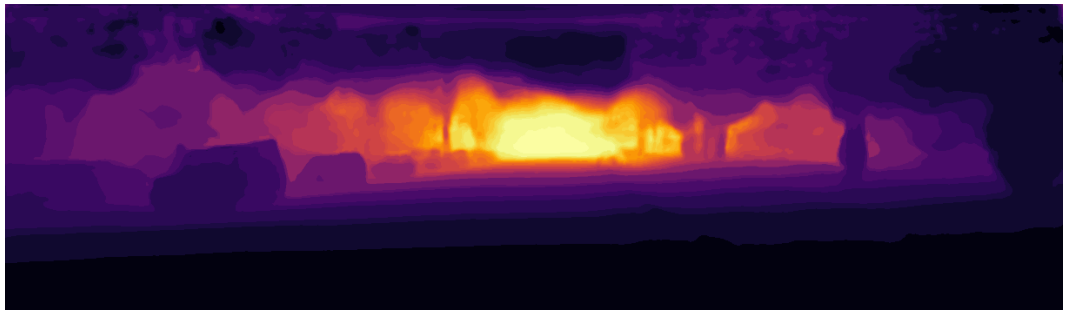


Original

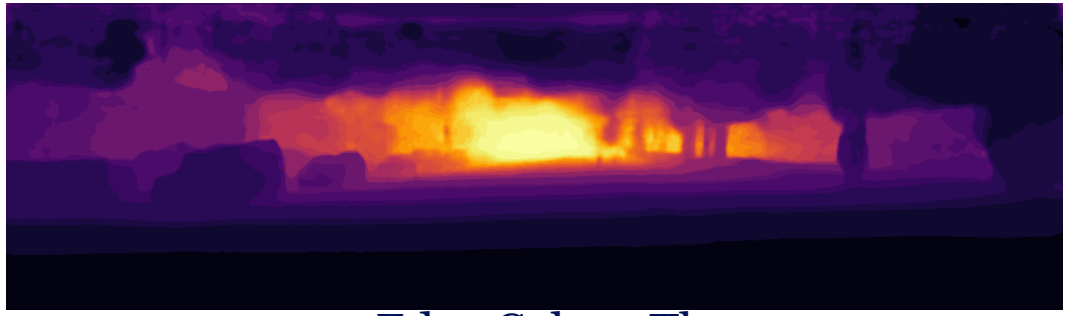
Sadly, no
Edge
Detection ☹



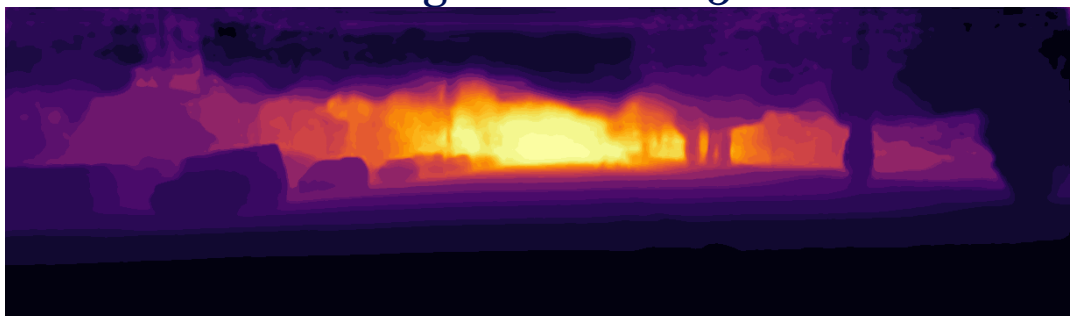
Edge Grey Th 5



Edge Colour Th 50

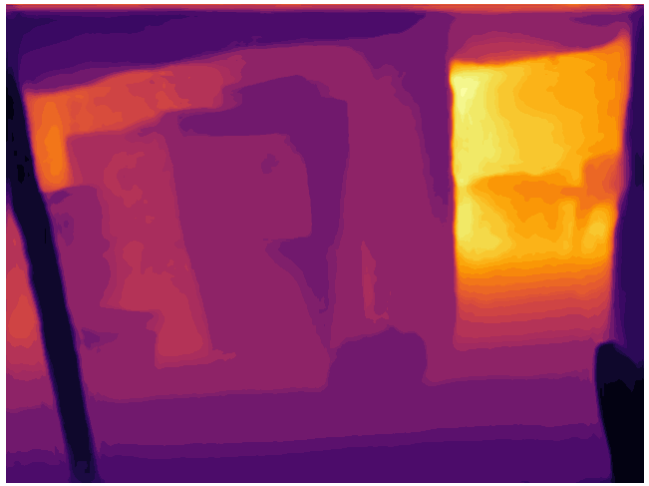


Edge Colour Th 5

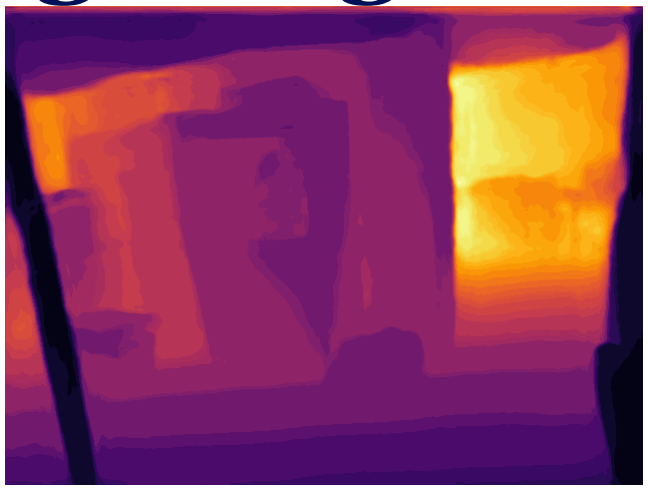


Edge Colour Th 50

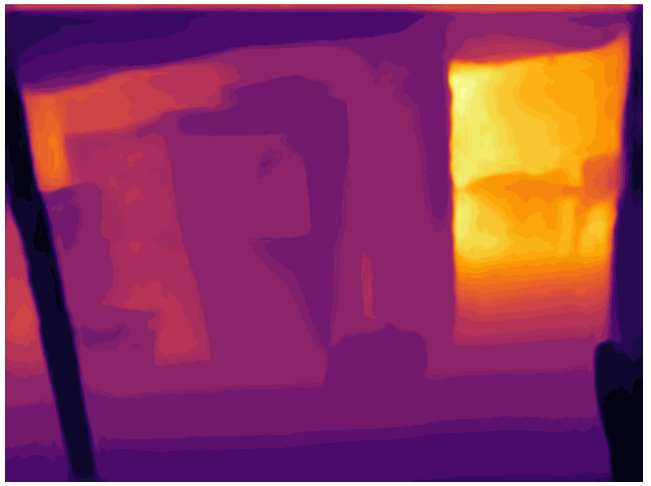
GLPN With Image Augmentation NYU



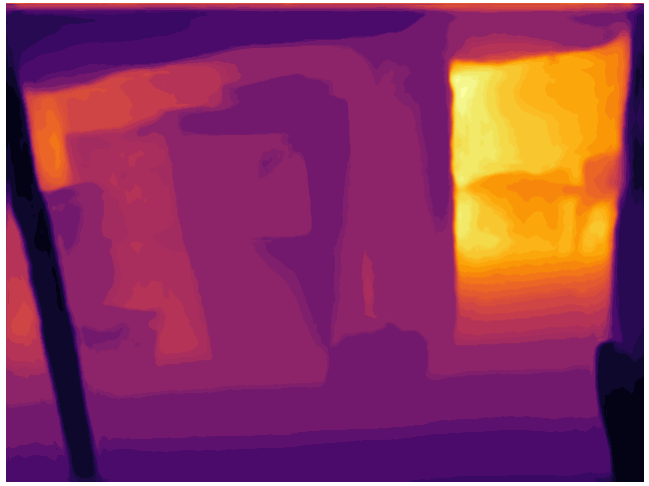
Original



Greyscale

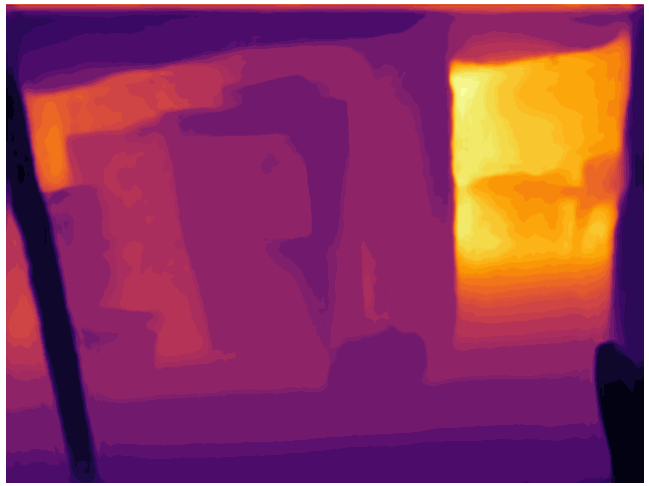


Desaturated

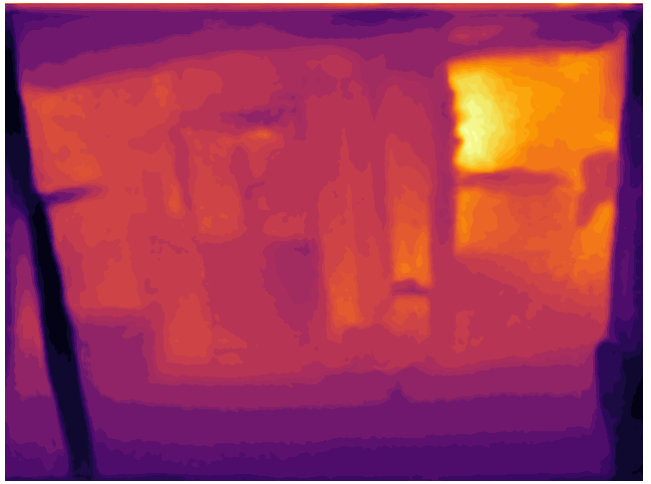


Oversaturated

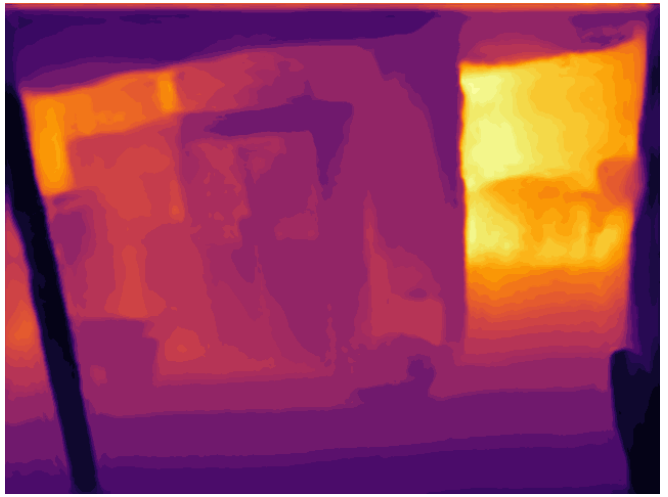
GLPN With Image Augmentation NYU



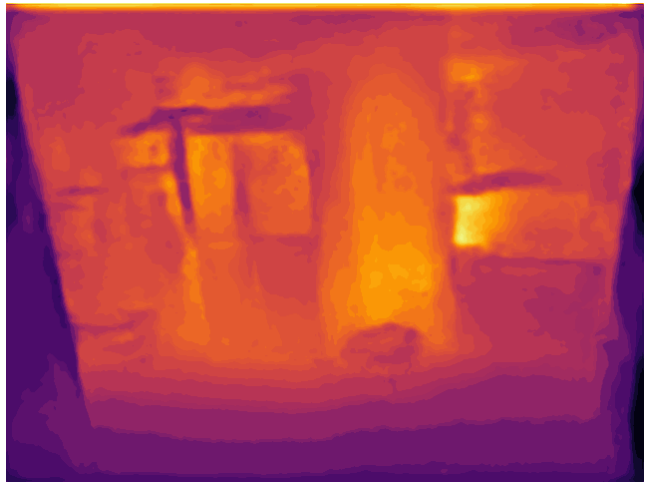
Original



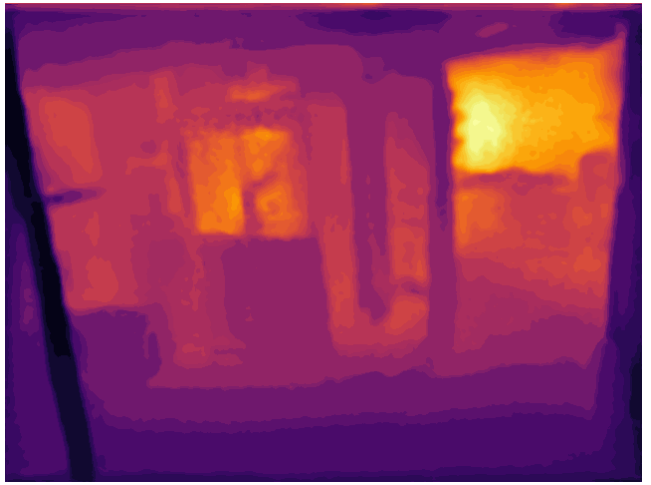
Edge Colour Th 5



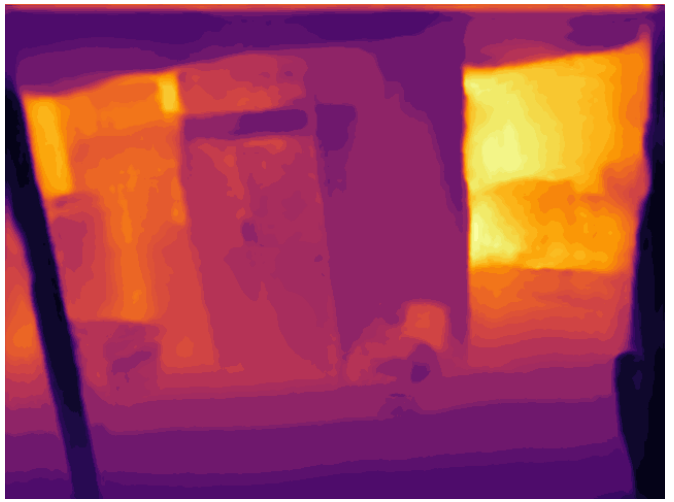
Edge Colour Th 50



Edge Detection



Edge Grey Th 5



Edge Grey Th 50

NeWCFs with Image augmentation

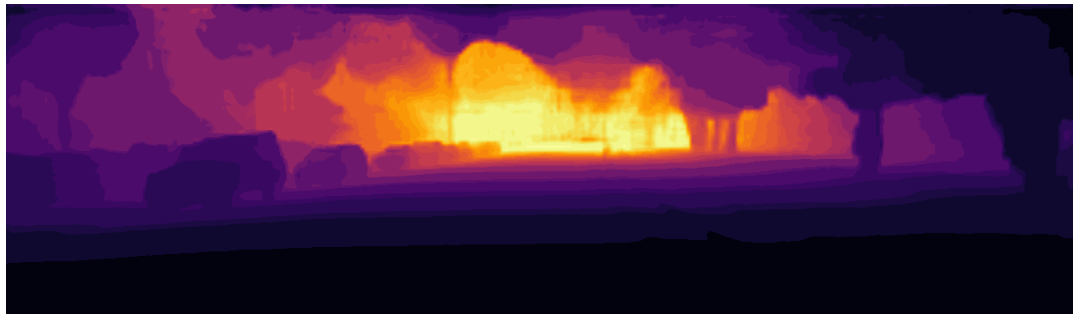
Preprocessing Type	silog	log10	rmse	rmse_log	abs_rel	sq_rel	d1	d2	d3
Original	5.6374	0.0178	1.9663	0.0629	0.0405	0.1192	0.9860	0.9980	0.9996
Greyscale	8.6504	0.0267	3.1456	0.0983	0.0580	0.2672	0.9530	0.9927	0.9981
Desaturated	6.0701	0.0192	2.1392	0.0679	0.0433	0.1384	0.9819	0.9974	0.9994
Oversaturated	6.1937	0.0190	2.1637	0.0683	0.0430	0.1347	0.9834	0.9976	0.9994
Edge Detection	35.2624	0.1729	11.8655	0.5308	0.2874	3.8204	0.4354	0.6621	0.7873
Edge Detection and Colour Threshold 5	17.1113	0.0789	6.9600	0.2446	0.1553	1.2788	0.7189	0.9128	0.9753
Edge Detection and Colour Threshold 50	7.4078	0.0289	3.0281	0.0935	0.0631	0.2468	0.9619	0.9957	0.9991
Edge Detection and Greyscale Threshold 5	24.4599	0.1175	9.2577	0.3607	0.2161	2.3049	0.5735	0.7977	0.9098
Edge Detection and Greyscale Threshold 50	11.6170	0.0430	4.6130	0.1475	0.0896	0.5529	0.8999	0.9762	0.9947

(a) Results of the experiments run on the KITTI Eigen Dataset

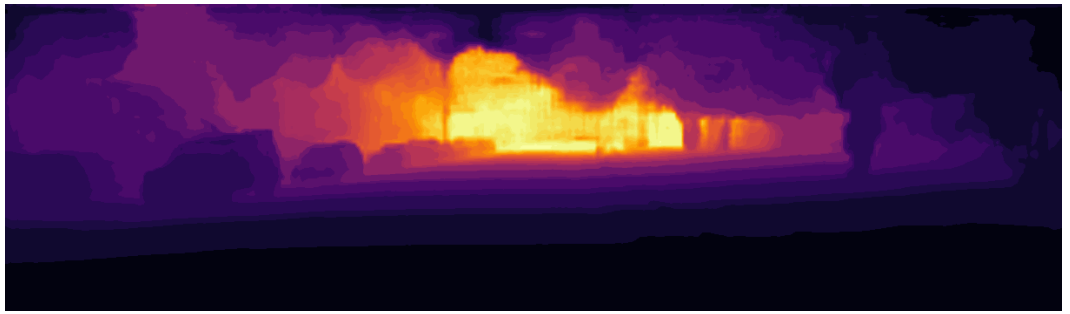
Preprocessing Type	silog	log10	rmse	rmse_log	abs_rel	sq_rel	d1	d2	d3
Original	3.8978	0.0150	0.1722	0.0474	0.0343	0.0078	0.9961	0.9994	0.9998
Greyscale	5.7664	0.0249	0.2910	0.0749	0.0555	0.0238	0.9880	0.9969	0.9989
Desaturated	4.0876	0.0166	0.1889	0.0511	0.0379	0.0093	0.9958	0.9993	0.9998
Oversaturated	4.0847	0.0159	0.1788	0.0498	0.0363	0.0084	0.9958	0.9993	0.9998
Edge Detection	22.3669	0.1298	1.3351	0.3580	0.2430	0.3904	0.4553	0.7855	0.9219
Edge Detection and Colour Threshold 5	10.9872	0.0480	0.6336	0.1475	0.1008	0.0953	0.8978	0.9779	0.9893
Edge Detection and Colour Threshold 50	4.3086	0.0171	0.1947	0.0533	0.0390	0.0101	0.9950	0.9991	0.9997
Edge Detection and Greyscale Threshold 5	14.0475	0.0690	0.8324	0.2015	0.1403	0.1614	0.7569	0.9560	0.9815
Edge Detection and Greyscale Threshold 50	6.2032	0.0274	0.3184	0.0818	0.0608	0.0285	0.9835	0.9963	0.9986

(b) Results of the experiments run on the NYU-Depth V2 Dataset

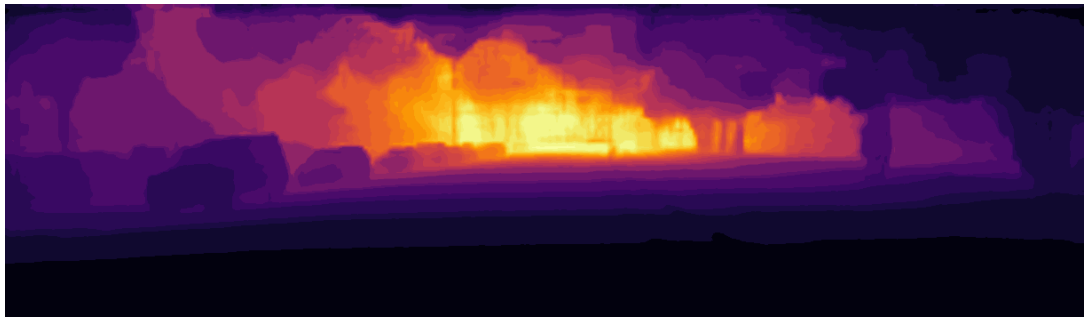
GLPN With Image Augmentation KITTI



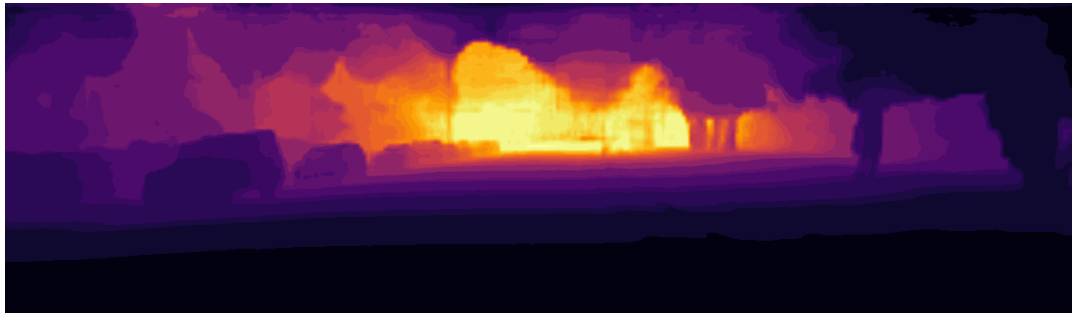
Original



Greyscale

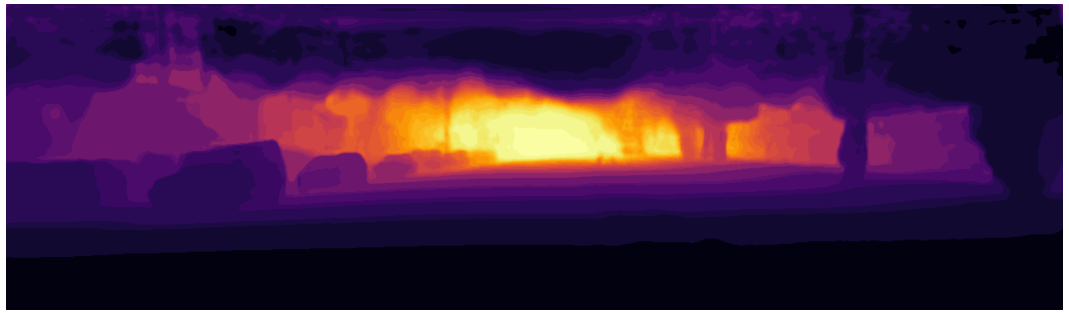


Oversaturated

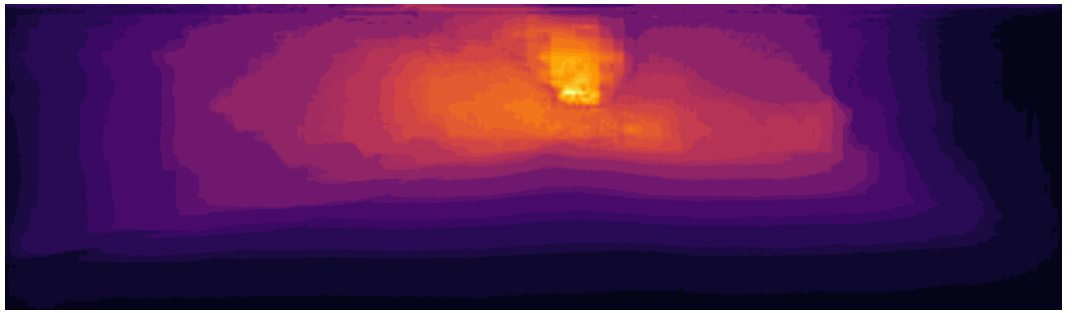


Desaturated

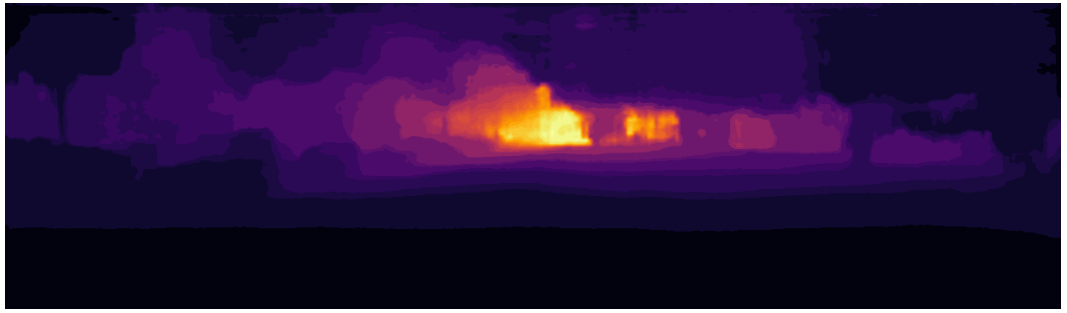
GLPN With Image Augmentation KITTI



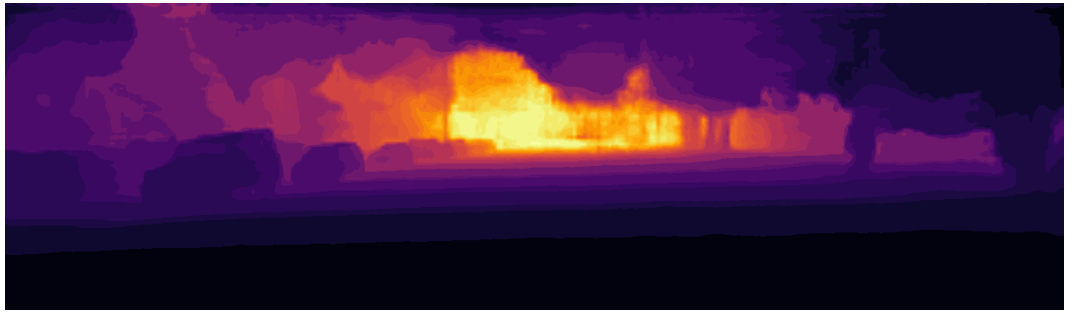
Original



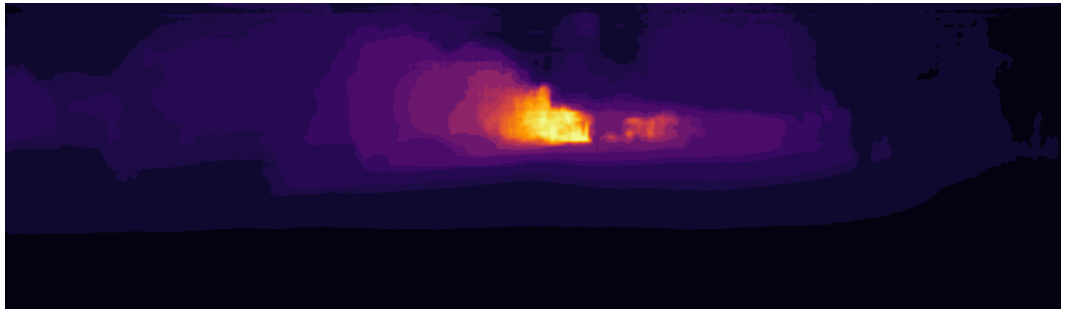
Edge Detection



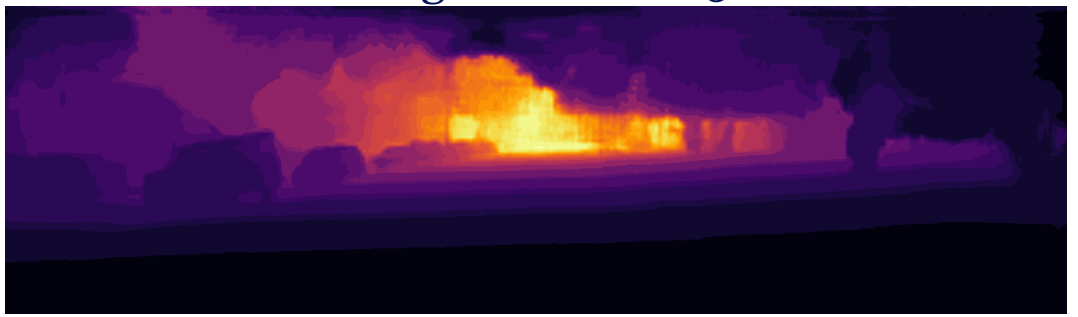
Edge Colour Th 5



Edge Colour Th 50

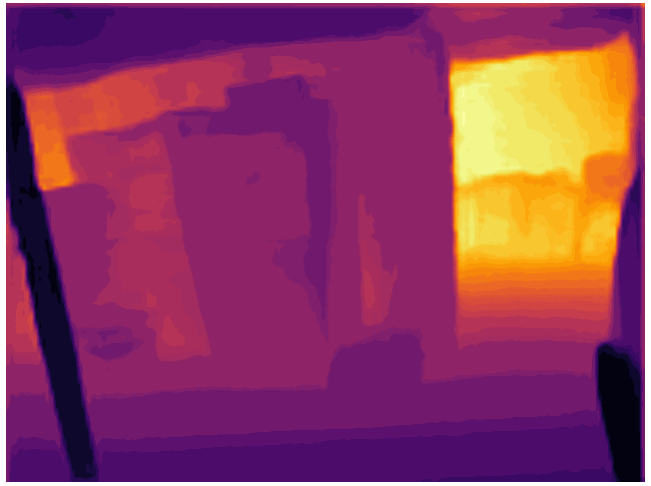


Edge Grey Th 5

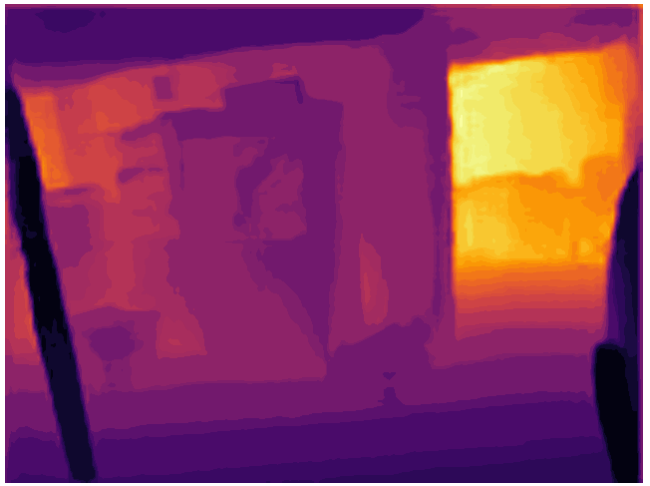


Edge Colour Th 50

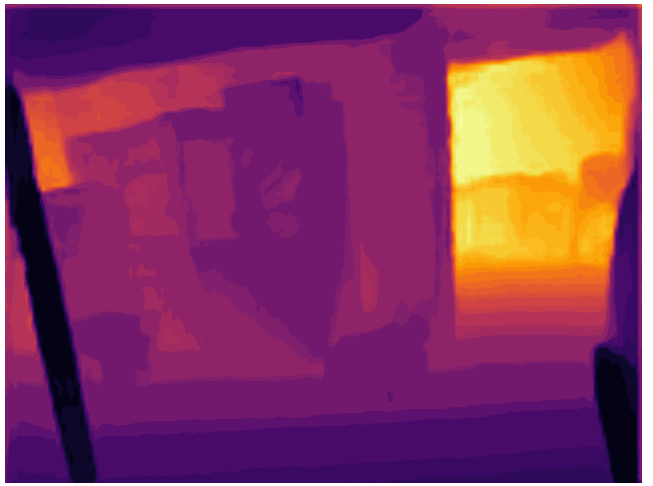
GLPN With Image Augmentation NYU



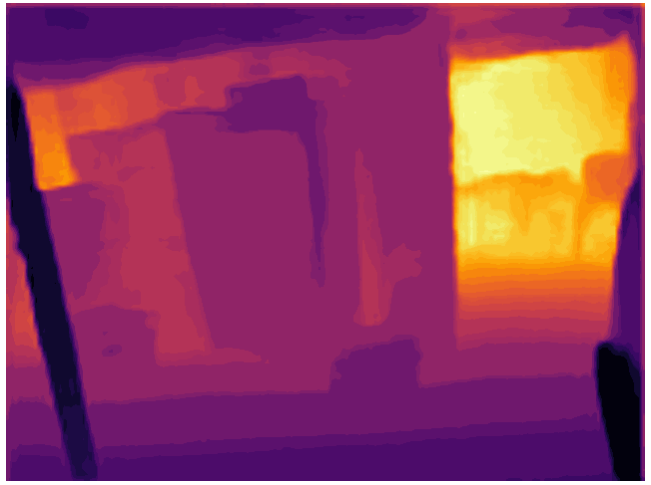
Original



Greyscale

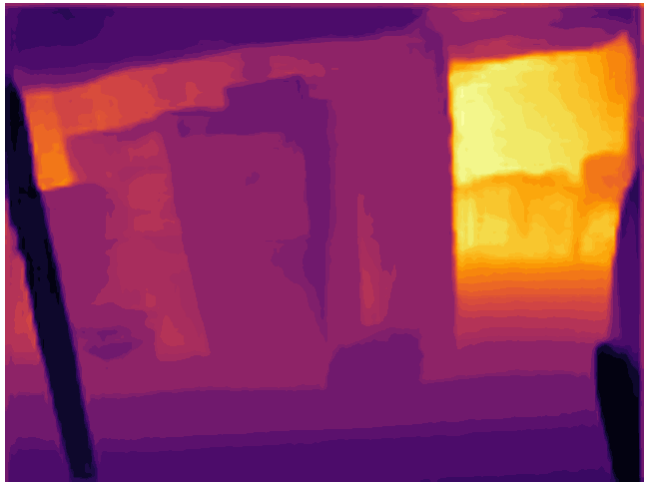


Desaturated

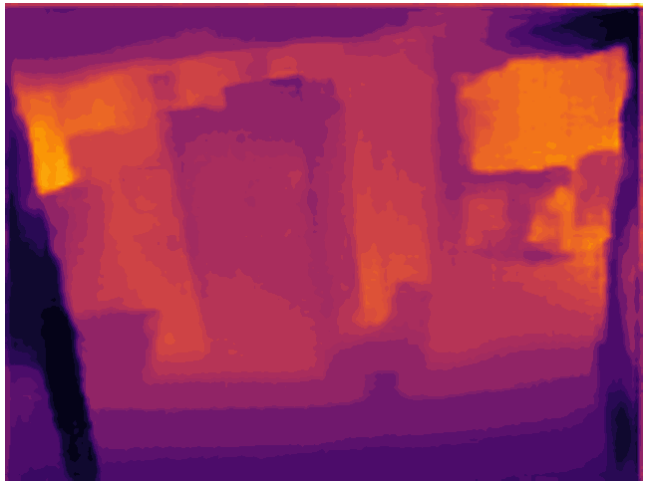


Oversaturated

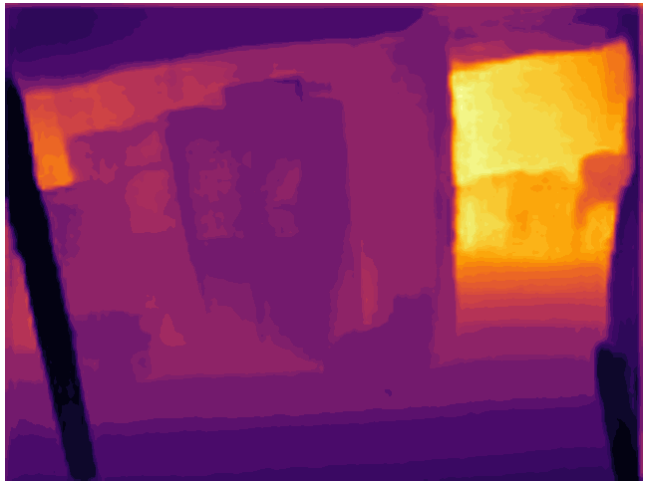
GLPN With Image Augmentation KITTI



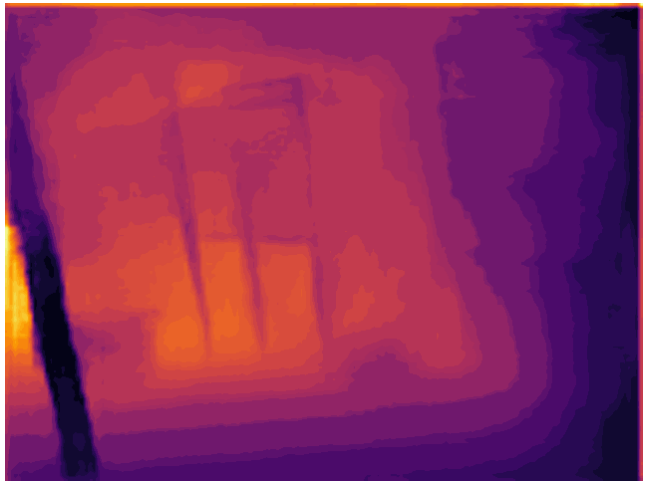
Original



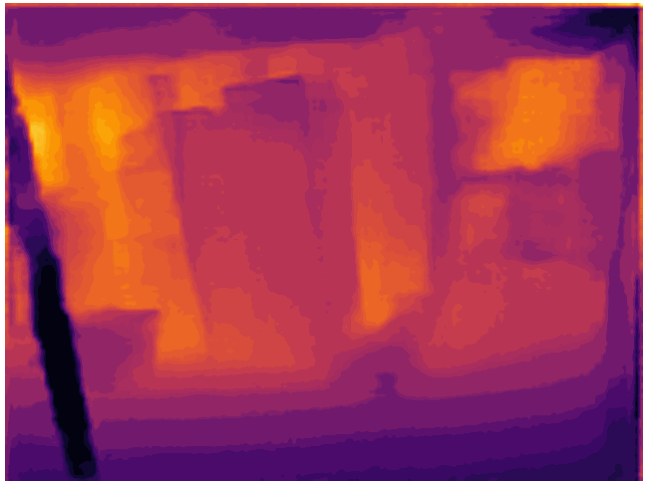
Edge Colour Th 5



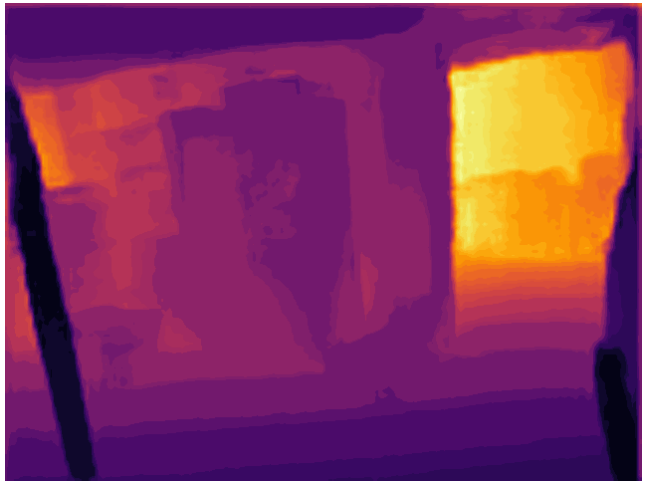
Edge Colour Th 50



Edge Detection



Edge Grey Th 5



Edge Grey Th 50

Conclusion

- Monocular-Depth Estimation extracts depth information from RGB images
- Compared 2 state-of-the-art methods: GLPN and NeWCFs
- Messed up the experiments so the results were not decisive
- Applied different image augmentation algorithms
- Changing the Colour did not change much
- Pure edge detection made the depth estimation worse
- Adding an image improved it

Title closure

Universiteit
Leiden
The Netherlands