Color-Light Multi Cascade Network for Single Image Depth Prediction on One Perspective Artifact Images

Aufaclav Zatu Kusuma Frisky[1,2], Simon Brenner[1], Sebastian Zambanini[1] and Robert Sablatnig[1]
[1]*Computer Vision Lab, Institute of Visual Computing and Human-Centered Technology, Faculty of Informatik, TU Wien, Austria*
[2]*Electronics and Instrumentations Lab, Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia*

Keywords: Single Image, Depth Prediction, Color-light, Multi Cascade, One-side Perspective, State-of-the-Art, Roman Coins, Temple Relief.

Abstract: Different color material and extreme lighting change pose a problem for single image depth prediction on archeological artifacts. These conditions can lead to misprediction on the surface of the foreground depth reconstruction. We propose a new method, the Color-Light Multi-Cascade Network, to overcome single image depth prediction limitations under these influences. Two feature extractions based on Multi-Cascade Networks (MCNet) are trained to deal with light and color problems individually for this new approach. By concatenating both of the features, we create a new architecture capable of reducing both color and light problems. Three datasets are used to evaluate the method with respect to color and lighting variations. Our experiments show that the individual Color-MCNet can improve the performance in the presence of color variations and fails to handle extreme light changes; the Light-MCNet, on the other hand, shows consistent results under changing lighting conditions but lacks detail. When joining the feature maps of Color-MCNet and Light-MCNet, we obtain a detailed surface both in the presence of different material colors in relief images, and under different lighting conditions. These results prove that our networks outperform state-of-the-art in limited number dataset. Finally, we also evaluate our joined network on the NYU Depth V2 Dataset to compare it with other state-of-the-art methods and obtain comparable performance.

1 INTRODUCTION

3D scanners are widely used these days for digital archiving of objects of cultural heritage (Georgopoulos et al., 2010); however, 3D scanning of scenes and multiple objects is time-consuming. The use of these scanners is subject to high maintenance configurations, such as the light, distance, and the number of scans (Georgopoulos et al., 2010). The lack of energy sources at sites that are difficult to access exacerbates this problem. If scanning using high precision scanners is too expensive, archaeologists search for time-efficient and robust alternatives (Frisky et al., 2020). In practical terms, Single Image Reconstruction (SIR) is more time-efficient when opposed to Structure-from-Motion and structure light scanning techniques, as 3D models can be obtained efficiently only using a single image. Distance prediction, referred to as Single Image Depth Prediction (SIDP), is one of the crucial steps in the reconstruction phase that uses in SIR (other than depth to 3D transfer), that decides the final product's quality (Ming et al., 2021).

Problems can arise in SIDP due to material object colors and extreme lighting conditions. The material color poses problems in two different scenarios that usually appear in real-life depth prediction (Frisky et al., 2021c). In the first scenario, surfaces at different depths appear in the same color, leading to misprediction at depth discontinuities. In the second scenario, different surface colors appear in regions of equal depth, leading to noise in the depth predictions.

Light conditions change the appearance of the artifact in the image. This phenomenon is also present in outdoor scenes and thus appears in NYU Depth V2 (Silberman et al., 2012). To the best of our knowledge, the handling of extreme light conditions is not addressed in a specific manner in previous SIDP research (Frisky et al., 2021b). An investigation in Frisky et. al (Frisky et al., 2021b) work shows that the state of the arts cannot handle the extreme change

Frisky, A., Brenner, S., Zambanini, S. and Sablatnig, R.
Color-Light Multi Cascade Network for Single Image Depth Prediction on One Perspective Artifact Images.
DOI: 10.5220/0010801100003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 909-916
ISBN: 978-989-758-555-5; ISSN: 2184-4321

of light. It inspired us to create a new method that is robust to both color and lighting variations. In specific areas, such as cultural heritage applications, most datasets contain a limited number of samples (Brenner et al., 2018; Frisky et al., 2021a). This condition also motivates a creation of a system that performs well on small datasets.

The contributions of this paper are summarized as follows:

1. We improve the architecture of the previous color robust network (Color-MCNet) by changing the number of dimensions of the transferred feature map.
2. We create a new Light-MCNet architecture using input from Intrinsic Image Decomposition to handle extreme lighting variations.
3. We create a new a combined architecture by joining the color- and Light-MCNet feature maps, thereby improving both extreme lighting and material color robustness of features in the final output. Our system outperforms state-of-the-art methods in small datasets.

2 RELATED WORK

Historically, single-image 3D reconstruction has been approached via shape-from-shading (Ruo Zhang et al., 1999). However, the pure shape from shading methods make use of only a single depth cue and are sensitive to color variations and depth discontinuities (Ming et al., 2021). Saxena et al. (Saxena et al., 2009) estimated depth from a single image by training a Markov Random Field on local and global image features. Oswald et al. (Oswald et al., 2012) improved the performance using interactive user input with the same depth estimation problem. In archaeology, the reconstruction of cultural objects using 2D images is used because of its flexibility and efficiency (Frisky et al., 2020). Pollefeys et al. (Pollefeys et al., 2001) present an approach that obtains virtual models. Regarding the color in archeology artifacts, two sub-problems need to be solved (Frisky et al., 2021c). First, an exemplary system needs to be able to predict the depth in the presence of different surface colors, as they appear on relief surfaces. Second, the system needs to reconstruct surfaces with different depths but similar material colors. For most artifacts, foreground and background (e.g. relief and wall) are made of the same material.

In order to predict the depth and the shape, most methods use RGB color as an input. In recent work, Pan et al. (Pan et al., 2018) reconstruct a Borobudur relief using single image reconstruction based on the multi-depth approach from Eigen and Fergus (Eigen and Fergus, 2015). However, the data used in their paper does not involve different material colors. For the requirement to solve these problems in the archaeological area, Frisky et al. provide a Registered Relief Depth (RRD) dataset consisting of RGB images and its corresponding depth on outdoor Borobudur (Frisky et al., 2021a), and Prambanan reliefs (Frisky et al., 2021c). Frisky et al. also propose a new method called MCCNet (Frisky et al., 2021c) that uses the cascaded color spaces with weight transfer to solve both mentioned color problems.

Varying lighting situations pose another problem for SIDP, which can affect performance and are difficult to control in natural environments. To the best of our knowledge, no specific research available in monocular depth reconstruction mentions this problem. Most recent research uses available datasets such as an indoor NYU Depth Dataset, that only use natural light (Song et al., 2021). However, no specific experiment shows that it is robust to extreme changes in lighting direction. Datasets of imaged under varying lighting conditions, together with a corresponding depth map, are commonly used in photometric stereo research (Brenner et al., 2018). Brenner et al. created a dataset of ancient roman coins, each illuminated from 54 directions (Brenner et al., 2018). It is believed that coin surface are similar to wall-reliefs in that they are sufficiently modelled in 2.5D representations (Frisky et al., 2021b). Together with a depth map created via photometric stereo, this Roman Coin dataset can be used as ground truth for evaluating the robustness of SIDP approaches with respect to lighting direction.

3 PROPOSED METHOD

In this work we propose a single image depth prediction method that is robust against material color and extreme lighting variations. Furthermore, a single network that can extract both properties (robust to color and light difference) with low computational time is needed.

First, our work utilizes an improved version of the previous cascaded network (Color-MCNet) (Frisky et al., 2021c). Second, we add a new network robust to lighting variations, called Light-MCNet. Then, a feature joining mechanism to produce the final result. An overview of the proposed method (Color-Light-MCNet) is given in Figure 1. In the following, the building blocks of the approach are described in detail.

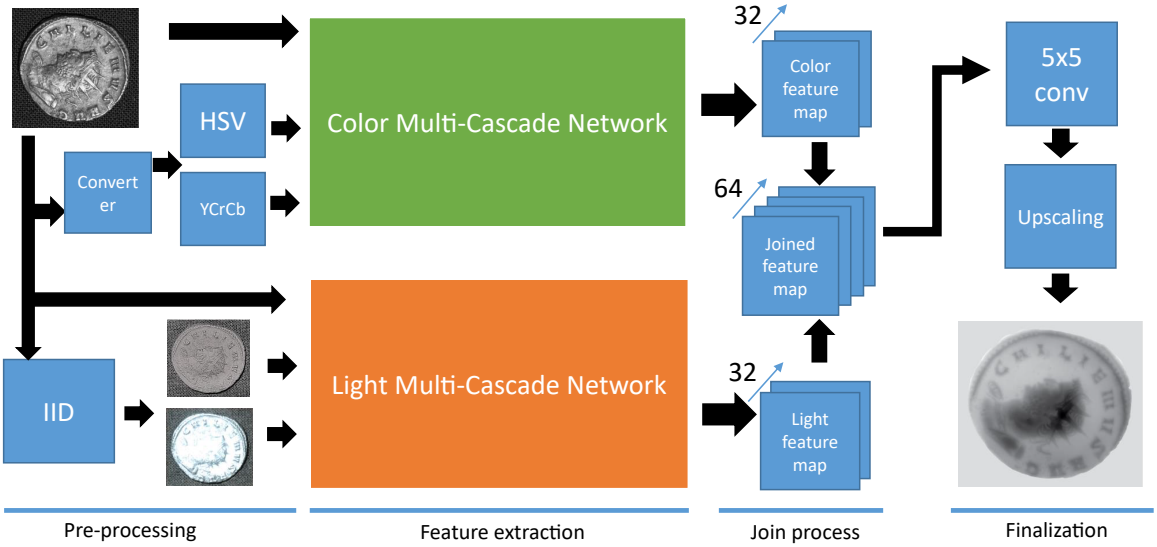

Figure 1: Architecture of the proposed Color-Light-MCNet. IID is Intrinsic Image Decomposition. The numbered arrows beside 32 and 64 indicate the dimensionality of the feature maps.

3.1 Color-MCNet

In the Color-MCNet section (see Figure 2, we use a modified version of the MCCNet architecture from Frisky et al. (Frisky et al., 2021c), which is designed for SIDP robust to surface color variations and is reused in this work to obtain color robust features. The input image is successively presented to the network in RGB, YCrCb and HSV color spaces, where features from each stage are concatenated with features from the previous stage. Our networks make use of the sub-architectures illustrated in Figure 4. In sub-architecture 1a, we use a 9x9 convolution network, a stride of 2 and 2x2 pooling. This configuration makes the feature map size a quarter compared to the input size, and a 3x3 convolution is applied afterwards. In sub-architecture 2a, similar to 1a, we used a 9x9 convolution network, a stride of 2 and 2x2 pooling. In sub-architecture 2b, the feature map of the previous cascade level is concatenated to the current feature map, and in 2c, similar to 1b, a 3x3 convolution is applied. The weight transfer in this network is applied from sub-architecture 1a to 2a and sub-architecture 1b to 2c. While the transferred feature maps in the original architecture were 1-dimensional (Frisky et al., 2021c), now a 32-dimensional feature map is transferred in order to allow more information to flow from one stage of the cascade to the next.

3.2 Light-MCNet

In the Light-MCNet section (see Figure 3), we change the input and parameters of Color-MCNet: the RGB

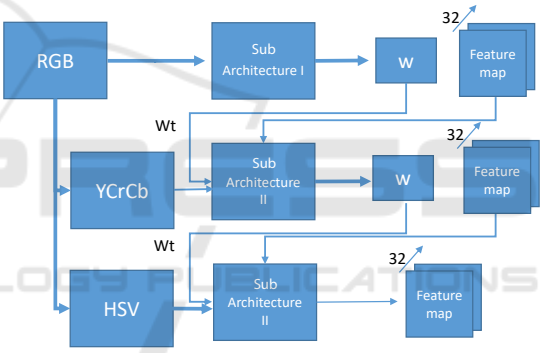

Figure 2: Three color-spaces (RGB,YCrCb,HSV) Color-MCNet feature extraction. The input image is successively presented to the cascaded network in different color spaces. Feature maps and weights (w) are propagated to subsequent cascade levels. Wt denotes the weight transfer process. Sub Architectures 1 (initial feature map generation) and 2 (feature map generation with concatenation) are given in Figure 4.

image is decomposed using the unsupervised intrinsic image decomposition by Letry et al. (Lettry et al., 2018), in order to separate the lighting effects from the original surface reflectance (see Figure 5 for an example). In conjunction with the RGB image on the first level, these two images become an input to the network.

3.3 Color-Light-MCNet

The two architectures mentioned above, Color and Light-MCNet, aim to solve their specific problems, i.e., material color and extreme lighting problems, re-

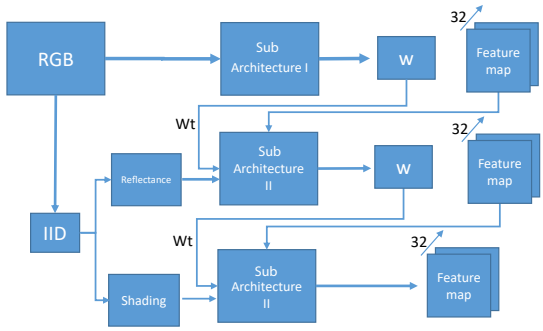

Figure 3: Light-MCNet feature extraction. On the first cascade level, the original RGB input image is presented to the network. On the subsquent levels, reflectance and shading components extracted by unsupervised intrinsic image decomposition (Lettry et al., 2018) are used as inputs. The rest of the network works analogous to the Color-MCNet (Figure 2).

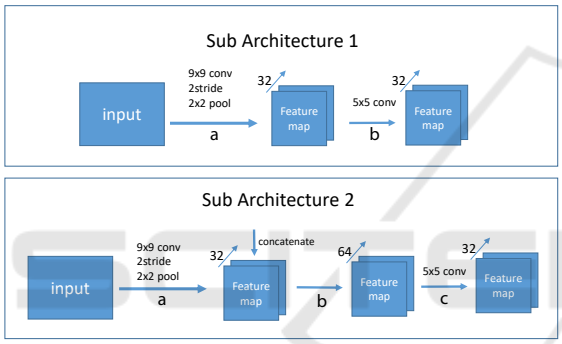

Figure 4: Sub architectures 1 and 2 that are used in Color-MCNet and Light-MCNet.

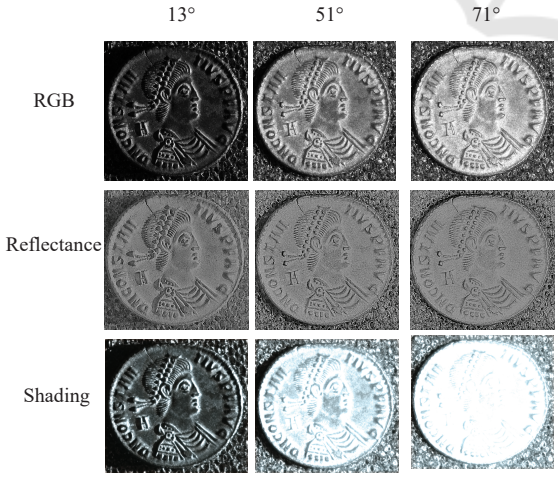

Figure 5: Example results of the intrinsic image decomposition for different light elevation angles (Roman Coin dataset).

spectively. However, a single architecture that can address both problems simultaneously would be preferable. Thus, we added a concatenation mechanism on the output feature map in each architecture. The integration aims to combine the color and light robust feature maps from the two previous architectures. The combined features are passed to a 5x5 convolution into one feature dimension at the end of this network. Finally, upscaling is carried out to return to the original resolution. The architecture of the proposed joint network is shown in Figure 1. The architecture consists of four parts: pre-processing, feature extraction, join process, and finalization. In preprocessing, conversion into different color-space is performed for Color-MCNet, and Intrinsic image Decomposition (IID) is done for the Light-MCNet. In the feature extraction part, color and light robust feature maps are extracted, which are subsequently combined in the join process to a single feature map. Lastly, the finalization part converts the feature map into the final depth prediction.

4 EXPERIMENTS AND RESULTS

The performance of our architecture is individually tested for robustness to color and lighting variations, using the dedicated RRD Temple dataset and the Roman Coin dataset, respectively. For evaluating the performance in a mixed environment and comparison to state-of-the-art methods, the NYU Depth V2 Dataset is used. Additionally to evaluating our full network (Color-Light-MCNet), we also test the performance of its components (Color-MCNet and Light-MCNet) individually. As a primary error metric, we use the Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{|N|}\sum_{i=1}^{N} ||y_i - y_i^*||^2} \qquad (1)$$

where y_i is groundtruth depth, y_i^* is predicted depth and N is the number of test points.

4.1 Datasets

Two specialized datasets represent the two main challenges addressed in this paper: the RRD Temple dataset represents different materials in a temple relief, and the Roman Coin dataset represents different lighting conditions in the image. Both RRD temple and Roman coin datasets have a limited number of samples. This condition motivates the creation of a system that can perform well using a small number of data. Additionally, in order to test the general applicability of our approach and compare it with state of the art, we also use the public NYU Depth V2 dataset.

These three datasets are described in the following subsections.

First dataset is The Registered Relief Depth (RRD) Temple dataset, consists of two relief datasets acquired at two Indonesian temples: Prambanan and Borobudur. The RRD Temple Dataset is created by Frisky et al. (Frisky et al., 2021c; Frisky et al., 2021a) to accommodate the color difference problem created by different materials appearing on archaeological reliefs and different color by chemical reaction. From the 41 reliefs of the RRD Prambanan dataset, we use 21 for training and 20 for testing. In the RRD Borobudur dataset, 20 were used as training and 10 as test sets. In total, we obtain 41 training examples and 31 test examples.

Second dataset is the Roman coin, consists of 23 coins, 11 of which are imaged from both sides and 12 from one side only, resulting in a total of 34 coin sides. The dataset was originally created for photometric stereo reconstruction, using a PhaseOne IQ260 Achromatic camera and a light dome with 54 individually controlled LED light sources for illumination (Brenner et al., 2018).In the training phase, each image is paired with its corresponding depth map. From the 34 coin sides (each represented by 54 input pairs), 26 are used for training and eight for testing. All the grayscale images are converted into RGB before processing them in order to fit the network input.

The last dataset is the NYU Depth v2 that offers images and depth maps for various indoor scenes (Silberman et al., 2012). The dataset includes 120K training samples and 654 test samples, but we only use a 50K subset to train our network. This dataset is used to compare our results to the current state-of-the-art methods, as a majority of publications use it for evaluation.

4.1.1 Augmentation

In this work, we augment the training data of the three datasets in several ways (Eigen and Fergus, 2015):

- **Scaling:** Input and target images are scaled by $s \in [1, 1.5]$, and the depths are divided by s.
- **Rotation:** Input and target are rotated by $r \in [-5, 5]$ degrees.
- **Color:** Input values are multiplied globally by a random RGB value $c \in [0.8, 1.2]^3$
- **Flips:** Input and target are horizontally flipped with 0.5 probability

4.2 Configuration

We test three different architectures: the individual Color-MCNet and Light-MCNet and the combined Color-Light-MCNet network. Both Color-MCNet and Light-MCNet use three different inputs on a three-level multi cascade network. As shown in Figures 2 and 3 both networks output 32 dimensional feature maps. In order to obtain the final depth estimations, these feature maps are passed to an additional 5x5 convolution and upscaling network, analogous to the finalization stage of the Color-Light-MCNet shown in Figure 1. For our experiments, we train and test the Color-MCNet, Light-MCNet and Color-Light-MCNet on all three datasets from scratch for 100 epochs with a learning rate of 0.001. Additionally, AdaBins (Bhat et al., 2020) is trained and tested on the RRD Temple dataset and the Roman Coin dataset for reference; the performance of AdaBins on the NYU Depth V2 Dataset is given by the authors (Bhat et al., 2020).

4.3 Results

The three datasets used in this work represent different problems: the RRD Temple dataset represents color problems, the Roman Coin dataset represents lighting problems, and the NYU Depth V2 dataset represents common indoor scenes and allows a comparison to state of the art. We thus evaluate our three architectures on these three datasets in order to assess their performance with respect to each of their specific problems. Results, including a comparison to AdaBins, are shown in TABLE 1.

On the RRD Temple dataset, the Color-MCNet performs better than the Light-MCNet. The difference in the material in the Prambanan temple and the yellowing color in the Borobudur temple can be appropriately resolved (see Figure 6). Compared to the two networks, color-Light-MCNet outperformed all implemented networks, including the AdaBins network. In Figure 6, it can be seen that Color-MCNet and Color-Light-MCNet perform well to different materials in relief. On the other hand, Light-MCNet produces low detail depth, and AdaBins results exhibit erroneous depth differences caused by different materials.

Regarding the performance on the Roman Coin dataset, the Color-MCNet architecture cannot resolve the influence of different lighting angles (see the second row of Figure 7). Within the dataset, this method performs most adequately with a 51° light elevation angle; this suggests that the Color-MCNet architecture needs a specifics light condition for maximum performance. For the Light-MCNet, we cannot observe significantly better results with respect to absolute RMS errors; for the input images with 51° elevation angle, it performs even worse than the Color-

Table 1: Results of the proposed method on three databases (RMSE in mm, lower better). For the Roman Coin dataset, results are grouped with respect to different light elevation angles.

	RRD Temple	Roman Coin					NYU Depth V2
		13°	32°	51°	71°	82°	
Color-MCNet	2.53	3.22	1.70	1.06	1.24	1.18	0.557
Light-MCNet	4.43	1.34	1.28	1.25	1.21	1.18	0.720
Adabins (Bhat et al., 2020)	2.43	1.14	0.82	0.95	1.92	2.04	**0.364**
Color-Light MCNet	**2.32**	**0.89**	**0.75**	**0.71**	**0.68**	**0.72**	0.376

MCNet. However, the results of this architecture are stable and are not impacted by different angles of incident light. This can also be observed in Figure 7: the architecture is robust to incident light direction, but the results lack detail.

AdaBins performs similar to Color-MCNet, but it produces better details on the results. The depth results of the Color-Light-MCNet are rich in detail and robust to varying lighting conditions; even in extreme light angles (13° and 81°, where the state of the arts fails to produce consistent outputs (see the two bottom rows of Figure 7 for example). Again, Color-Light-MCNet outperforms the other tested methods with respect to RMSE. Given the limited size of the dataset, these results also suggest that our method is especially useful in such situations.

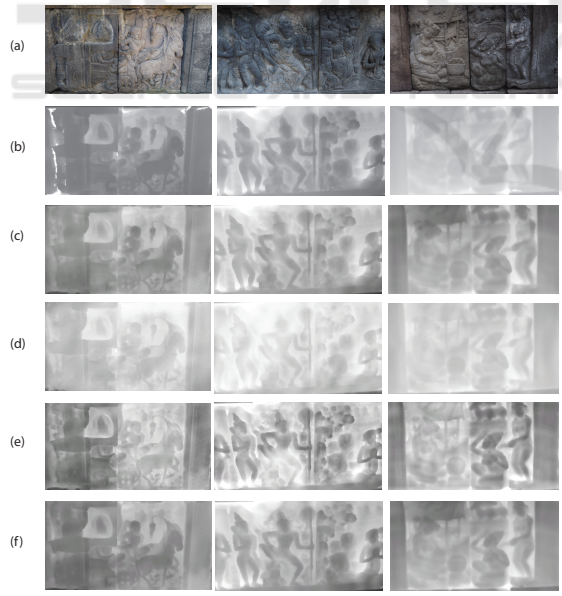

Figure 6: Example results for the RRD Temple datasets. a: RGB image (input), b: ground truth, c: Color-MCNet, d: Light-MCNet, e: AdaBins, f: Color-Light-MCNet.

Using the NYU Depth V2 dataset, we test the performance of our methods on a public dataset and compare it to multiple states of the art methods. In Table 1, it can be seen that Color-MCNet and Light-MCNet perform worst, while AdaBins and Color-Light-MCNet show similar results. TABLE 2 shows a comparison of Color-Light-MCNet to state of the art methods on the NYU Depth V2 dataset using several error metrics, such as RMSE (Equation 1), threshold (Equation 2), RMSE log (Equation 3), and absolute Relative Difference (abs.REL) (Equation 4). The state of the art results given in TABLE 2 are taken from available scoreboards (Bhat et al., 2020). It can be observed that on the NYU Depth V2 dataset, our proposed Color-Light-MCNet achieves competitive results with state of the art in all comparison metrics.

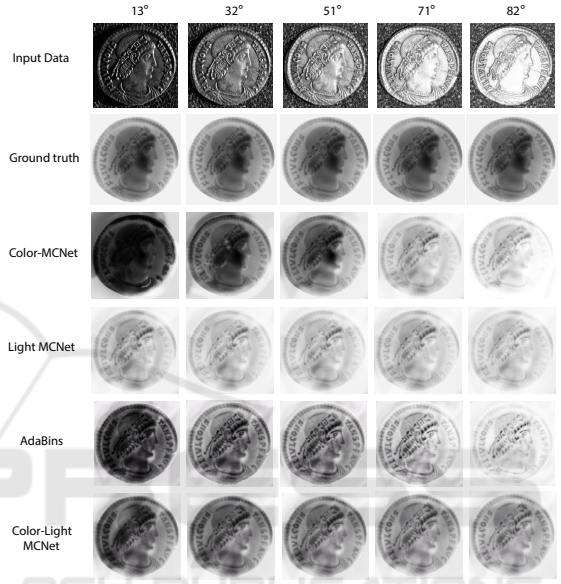

Figure 7: Results of the proposed method on the Roman Coin datasets. Columns represent different elevation angles of incident light. The first row shows the input images, the subsequent rows show the results obtained from different architectures.

$$\% \; of \; y_i \; s.t. \; max(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}) = \delta < thr \qquad (2)$$

$$\text{RMSE log} = \sqrt{\frac{1}{|N|}\sum_{i=1}^{N} ||log\, y_i - log\, y_i^*||^2} \qquad (3)$$

Table 2: Results of the proposed method on NYU Depth V2 compared with other state-of-the-art methods.

	Higher better			Lower better		
	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	RMSE linear	RMSE log	abs. REL
(Eigen and Fergus, 2015)	77.10%	95.00%	98.80%	0.639	0.215	0.158
(Frisky et al., 2021c)	79.40%	95.50%	99.10%	0.598	0.202	0.145
(Lee et al., 2018)	81.50%	96.30%	99.10%	0.572	0.193	0.139
(Lee and Kim, 2019)	83.70%	97.10%	99.40%	0.538	0.180	0.131
(Lee et al., 2019)	88.50%	97.80%	99.40%	0.392	0.142	0.110
Proposed	92.17%	98.70%	99.50%	0.376	0.098	0.108
(Wu et al., 2019)	93.20%	98.90%	99.70%	0.382	**0.050**	0.115
(Bhat et al., 2020)	**90.30%**	**98.40%**	**99.70%**	**0.364**	0.088	**0.059**

$$\text{abs. REL} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - y_i^*|}{y_i} \tag{4}$$

5 CONCLUSION

In this work, the Color-Light-MCNet is proposed as a new approach for SIDP, specifically addressing depth mispredictions arising from variations in material color and lighting direction. Evaluations are performed with respect to three datasets: the robustness of the method to color and light variations is tested using the RRD Temple dataset and the Roan Coin dataset, respectively, while a comparison with state of the art in natural environments is performed using the public NYU Depth V2 dataset.

Prior to the full Color-Light-MCNet, we test its two main components individually: the Color-MCNet designed to produce features robust to surface color variations, and the Light-MCNet designed to produce features robust to lighting direction. The Color-MCNet performs well in the RRD Temple dataset but fails to resolve the influence of different lighting angles appearing in the Roman coin dataset; the method delivers acceptable results only for specific lighting directions.

The Light-MCNet introduces intrinsic image decomposition as a pre-processing step to separate the input images' lighting effects and surface reflectance. Together with the original RGB image, the decomposition results are presented as inputs to the cascade network.This approach proved largely invariant to lighting direction (Roman Coin dataset), but the results generally lack detail.

Finally, we combine the two feature maps obtained from Color-MCNet and Light-MCNet into a single stream to combine the strengths of both approaches in a single architecture. The resulting Color-Light-MCNet shows superior results on both the RRD Temple dataset and the Roman Coins dataset, where the results exhibit rich details and robustness to variations in surface color and lighting direction, respectively. For these datasets, our method clearly outperforms AdaBins (Bhat et al., 2020), the current state of the art SIDP method. The results make it evident of the superior performance of our method with small datasets. On the NYU Depth V2 dataset, Color-Light-MCNet could not outperform AdaBins but shows competitive performance with respect to another state of the art methods. Most of the images in this work were taken from a frontal view perspective of the artifact. In the future, more comprehensive research in SIDP on non-frontal views is needed.

ACKNOWLEDGMENT

This work is funded by a collaboration scheme between the Ministry of Research and Technology of the Republic of Indonesia and OeAD-GmbH within the Indonesian-Austrian Scholarship Program (IASP). This work is also supported by Type C grant from Electronics Instrumentation Lab, Universitas Gadjah Mada.

REFERENCES

Bhat, S. F., Alhashim, I., and Wonka, P. (2020). Adabins: Depth estimation using adaptive bins. *Arxiv*, abs/2011.14141.

Brenner, S., Zambanini, S., and Sablatnig, R. (2018). An investigation of optimal light source setups for photometric stereo reconstruction of historical coins. In *Eurographics Workshop on Graphics and Cultural Heritage*.

Eigen, D. and Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of*

IEEE International Conference on Computer Vision (ICCV), pages 2650–2658.

Frisky, A. Z. K., Fajri, A., Brenner, S., and Sablatnig, R. (2020). Acquisition evaluation on outdoor scanning for archaeological artifact digitalization. In Farinella, G. M., Radeva, P., and Braz, J., editors, *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2020, Volume 5: VISAPP, Valletta, Malta, February 27-29, 2020*, pages 792–799. SCITEPRESS.

Frisky, A. Z. K., Harjoko, A., Awaludin, L., Dharmawan, A., Augoestien, N. G., Candradewi, I., Hujja, R. M., Putranto, A., Hartono, T., Suhartono, Y., Zambanini, S., and Sablatnig, R. (2021a). Registered relief depth (rrd) borobudur dataset for single-frame depth prediction on one-side artifacts. *Data in Brief*, 35:106853.

Frisky, A. Z. K., Harjoko, A., Awaludin, L., Zambanini, S., and Sablatnig, R. (2021b). Investigation of single image depth prediction under different lighting conditions: A case study of ancient roman coins. *J. Comput. Cult. Herit.*, 14(4).

Frisky, A. Z. K., Putranto, A., Zambanini, S., and Sablatnig, R. (2021c). Mccnet: Multi-color cascade network with weight transfer for single image depth prediction on outdoor relief images. In Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G. M., Mei, T., Bertini, M., Escalante, H. J., and Vezzani, R., editors, *Pattern Recognition. ICPR International Workshops and Challenges*, pages 263–278, Cham. Springer International Publishing.

Georgopoulos, A., Ioannidis, C., and Valanis, A. (2010). Assessing the performance of a structured light scanner. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVIII:250–255.

Lee, J., Heo, M., Kim, K., and Kim, C. (2018). Single-image depth estimation based on fourier domain analysis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 330–339.

Lee, J. H., Han, M.-K., Ko, D. W., and Suh, I. H. (2019). From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*.

Lee, J.-H. and Kim, C.-S. (2019). Monocular depth estimation using relative depth maps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738.

Lettry, L., Vanhoey, K., and Van Gool, L. (2018). Unsupervised Deep Single-Image Intrinsic Decomposition using Illumination-Varying Image Sequences. *Computer Graphics Forum*, 37(10):409–419.

Ming, Y., Meng, X., Fan, C., and Yu, H. (2021). Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33.

Oswald, M. R., Töppe, E., and Cremers, D. (2012). Fast and globally optimal single view reconstruction of curved objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 534–541.

Pan, J., Li, L., Yamaguchi, H., Hasegawa, K., Thufail, F. I., Mantara, B., and Tanaka, S. (2018). 3D Reconstruction and Transparent Visualization of Indonesian Cultural Heritage from a Single Image. In *Eurographics Workshop on Graphics and Cultural Heritage*, pages 207–210. The Eurographics Association.

Pollefeys, M., Van Gool, L., Vergauwen, M., Cornelis, K., Verbiest, F., and Tops, J. (2001). Image-based 3d acquisition of archaeological heritage and applications. In *VAST 01: Proceedings of the 2001 conference on Virtual Reality, Archeology, and Cultural Heritage*, pages 255–262.

Ruo Zhang, Ping-Sing Tsai, Cryer, J. E., and Shah, M. (1999). Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706.

Saxena, A., Sun, M., and Ng, A. Y. (2009). Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *ECCV 2012*, pages 746–760, Berlin, Heidelberg. Springer Berlin Heidelberg.

Song, M., Lim, S., and Kim, W. (2021). Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1.

Wu, Y., Boominathan, V., Chen, H., Sankaranarayanan, A., and Veeraraghavan, A. (2019). Phasecam3d — learning phase masks for passive single view depth estimation. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12.