Uni MS-PS: a Multi-Scale Encoder Decoder Transformer for Universal Photometric Stereo

Clément Hardy, Yvain Quéau, David Tschumperlé

Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Uni MS-PS: a Multi-Scale Encoder Decoder Transformer for Universal Photometric Stereo

Clément Hardy[a,**], Yvain Quéau[a], David Tschumperlé[a]

[a]*Normandie Univ, UNICAEN, CNRS, ENSICAEN, GREYC laboratory, Caen, France*

ABSTRACT

Photometric Stereo (PS) addresses the challenge of reconstructing a three-dimensional (3D) representation of an object by estimating the 3D normals at all points on the object's surface. This is achieved through the analysis of at least three photographs, all taken from the same viewpoint but with distinct lighting conditions. This paper introduces a novel approach for Universal PS, i.e., when both the active lighting conditions and the ambient illumination are unknown. Our method employs a multi-scale encoder-decoder architecture based on Transformers that allows to accommodates images of any resolutions as well as varying number of input images. We are able to scale up to very high resolution images like 6000 pixels by 8000 pixels without losing performance and maintaining a decent memory footprint. Moreover, experiments on publicly available datasets establish that our proposed architecture improves the accuracy of the estimated normal field by a significant factor compared to state-of-the-art methods. Code and dataset available at: `https://clement-hardy.github.io/Uni-MS-PS/index.html`

1. Introduction

Photometric stereo (PS) is a technique for recovering surface normals of an object by capturing multiple images of it from the same perspective but under varying light conditions. For decades, traditional image processing methods have focused on the ideal Lambertian case with a controlled and parallel light beam as well as no ambient light [Woodham (1980)]. Howewer in practice most light effects on real-world objects deviate from Lambert's law, exhibiting complex effects such as specular components or translucency (e.g., transparent materials). On the other hand, the emergence of deep learning approaches has enabled significant advancements in managing more complex geometries and challenging objects that do not adhere to Lambert's law.

Three types of approachs are considered in the literature to adress the PS problem: calibrated, uncalibrated, and Universal methods. The difference between calibrated and uncalibrated methods lies in whether we know the light parameters (positions, intensities,...). Additionally, most of these methods (uncalibrated or calibrated) assume the ideal case of perfect directional lighting in a dark environment with no external light. Obtaining this ideal case in real life is challenging, requiring special equipment to capture images under such conditions. Universal methods overcome this limitation by reconstructing objects in any lighting conditions, thus largely simplifying the process from the end-user perspective. They simultaneously address two major challenges:

- reconstructing the normal map for non-Lambertian materials like specular ones;
- handling complex illumination conditions, including ambient illumination.

In our conference paper [Hardy et al. (2023)], we introduced a multi-scale approach to improve the performance of *calibrated* PS on challenging materials. In the present article, we extend this multi-scale approach to solve the Universal PS problem. We define a multi-scale architecture combined with an encoder-decoder architecture. The multi-scale architecture can process input images of any size without loss of performance, even when considering very high-resolution images, as presented in Figure 1. In this example, our algorithm takes 11

**Corresponding author:
e-mail: `clement.hardy@unicaen.fr` (Clément Hardy)

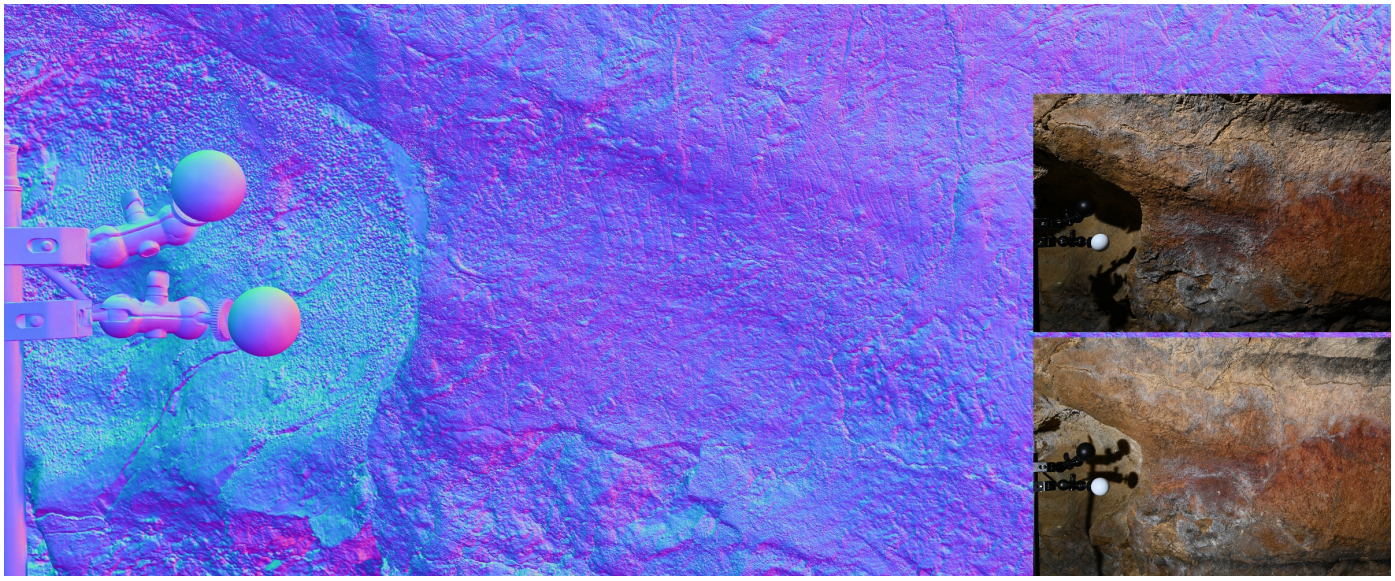


Fig. 1: Reconstruction of a normal map of size 6000×8000 of the Marsoulas cave by photometric stereo with the proposed method. Our method is able to both achieve accurate 3D-reconstruction and manage very high image resolution. Indeed, all details in the rock are reconstructed and the resolution of the RGB images is preserved. PS image credits: A. Laurent 2023 (INPT, UMR 5505 IRIT), C. Fritz and G. Tosello team (CREAP-E.Cartailhac), MSHS-T (UAR 3414).

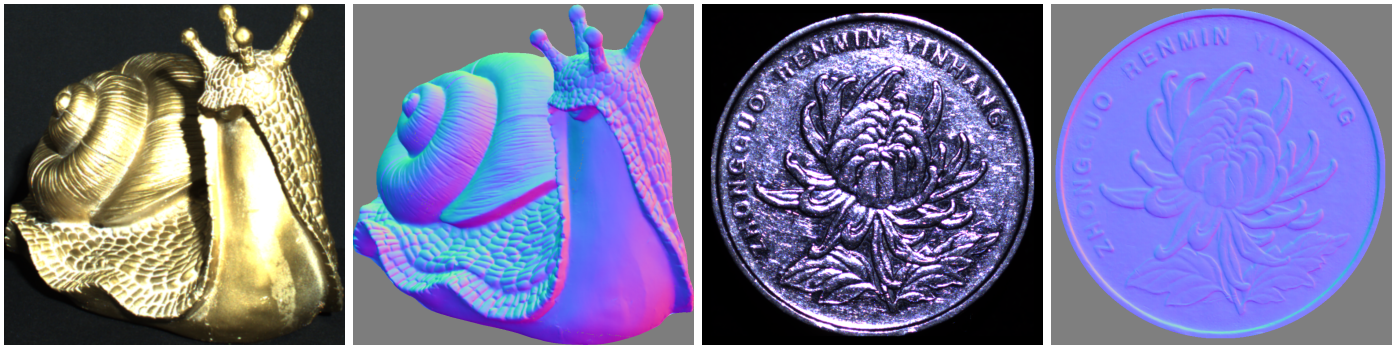


Fig. 2: Reconstruction of a normal map of a snail from Voynov et al. (2023) and a coin from Wang et al. (2023) by photometric stereo with our proposed method. We can see that our method manages very difficult materials such as metal.

images of size 6000×8000 as input, and we observe that our method recovers all details of the scene. Our method is also able to manage more difficult materials as illustrated in Figure 2. Our Universal method achieves state-of-the-art results in any environment or lighting conditions, including directional or non-parallel lighting beams, as demonstrated on real benchmarking datasets.

The main advantage of our proposed extension of Hardy et al. (2023) is the ability to solve the PS problem *in any light conditions*, without any prior knowledge of the environment (our previous method was only capable of handling calibrated PS in a controlled dark environment). The extension includes new two major features:

1. A Transformer-based approach is chosen for its effectiveness in Universal PS, instead of the CNN approach in Hardy et al. (2023).
2. A new training dataset is designed and synthetised to better cope with the Universal context.

The rest of this work is organised as follow. In Section 2, we present an overview of deep learning methods for photometric stereo. In Section 3, we describe our multi-scale Transformer method and our new training dataset. Finally, in Section 4, we present qualitative and quantitative results on many benchmark datasets and compare the performance of our method with state-of-the-art PS methods.

2. Related Work

Deep learning methods for PS proposed in the literature initially considered specific lighting conditions. Most of them, such as Chen et al. (2018, 2019), Ikehata (2018, 2022b), Yao et al. (2020) and Zheng et al. (2019), consider the light beam as parallel, which is not realistic in practice. Lichy et al. (2022) introduced methods for non-parallel light beams, but they still require images captured in a very controlled environment. All of these constraints limit their real-world applicability.

Calibrated PS — To solve the problem of calibrated photometric stereo, multiple learning-based methods have been introduced in the literature.

Santo et al. (2017) propose a first approach based entirely on a fully connected network. This approach assumes that the light directions are fixed and identical between the training and testing phases, which is its major drawback. Photometric stereo methods should be able to handle a different number of input images. Indeed, to make PS methods more suitable for experimentation, the neural network architecture should allow the use of an arbitrary number of images to avoid training a different neural network for each possible number of input images. Two main alternatives have been proposed in the literature to solve this problem:

• the first approach is to aggregate information from different images using a pooling layer, such as in the work of Chen et al. (2018, 2019), Hardy et al. (2023), Ju et al. (2021, 2020), Lichy et al. (2021) and Wang et al. (2020).
• the second approach is to project all observations corresponding to the same pixel location under different illuminations onto a fixed-size space, such as in the work of Ikehata (2018, 2022a), Li et al. (2019), Logothetis et al. (2021) and Zheng et al. (2019).

More recently, Ikehata (2022b) show the relevance of using Transformers and attention mechanisms in the context of calibrated PS with a small number of different lights.

Uncalibrated PS — Uncalibrated PS is a category of PS where the prior light information, such as its direction and intensity, is unknown. In the context of parallel light beams, a common practice is to use a first neural network to infer the missing light information for a standard calibrated PS neural network as presented in Chen et al. (2019). Then, a second neural network handles the problem as in the calibrated case. This approach has also been successfully applied to non-parallel light beams in Lichy et al. (2022). However, it is difficult to apply this approach to Universal PS with natural light/ambient light because the physics of natural light/ambient light is complex to model.
Another practice to solve uncalibrated PS is to use an inverse rendering based method [Li et al. (2023); Li and Li (2022); Kaya et al. (2021)]. Those methods optimize an image reconstruction loss (between the reconstructed images and the input images) to get the normal map, albedo...

Universal PS — Recently, Ikehata introduced the notion of Universal PS with the UNI-PS [Ikehata (2022c)] and SDM-UniPS methods [Ikehata (2023)]. These new methods solve the PS problem under unknown and arbitrary lighting conditions using a pure data-driven approach without complex prior light assumptions. This method is based on an encoder-decoder model, where the encoder extracts a global lighting context from a fixed 'canonical' resolution image. Images are resized if their size does not match the canonical resolution. The decoder takes as input the original resolution images and the output of the encoder, i.e., the global lighting context interpolated to the original resolution. This allows for inference on very high-resolution images, as the decoder processes the images pixel by pixel.

The idea behind using a global lighting context, rather than a global lighting model, is due to the spatially-varying light direction. Indeed, intensity could not be encoded by a few global values. Even though this method allows for processing high-resolution images by downsampling and inferring the normal map pixel by pixel, this technique decreases the performance of the reconstruction due to the loss of information during downsampling and the lack of spatial information during pixel-by-pixel inference.

To address this problem, Ikehata (2023) introduced a way to use all available information in a non-local way, even on very high-resolution images. The method is based on a scale-invariant spatial-light feature encoder, which allows for a fixed input size without resizing the images. The encoder splits an image into P^2 sub-images, where P is the size of the input of the model, by taking a single pixel every $P \times P$ pixels. It then extracts feature maps from these sub-images, which are eventually merged back to reconstruct one image.

During the encoding phase, spatial information is extracted using ConvNeXt layers [Liu et al. (2022)] and information over the light axis is extracted using Self-Attention Blocks [Lee et al. (2019)]. Afterwards, another pixel sampling strategy is used and several Transformer layers are applied in both the spatial and light dimensions. Finally, the normal map is inferred using two linear layers.

UNI-PS [Ikehata (2022c)] tends to infer blurry and inaccurate normal maps with a lack of detail, especially with high-resolution images. SDM-UniPS [Ikehata (2023)] produces better normal maps, but its performance decreases as object materials become more complex. In Ikehata (2023), Ikehata explains the importance of objects and materials diversity in the training dataset and so introduce a new dataset for training UNI-PS [Ikehata (2022c)]. Moreover, we have showed in Hardy et al. (2023) the more diverse and representative the training dataset is, the better the results are. To this end, we generate in the present paper a way more diverse, complete training dataset for Universal PS. Additionally, processing objects with complex materials, such as highly specular materials, requires both global and local information. We also showed in Hardy et al. (2023) that a multi-scale network architecture can extract both local and global information and gives thus gives excellent performance on non-Lambertian reflectance materials.

While UniPS allows for the processing of high-resolution images, it may be difficult to scale to very high resolutions without losing performance. Indeed, keeping only one pixel every P pixels is a problem if P is large (for instance if $P \geq 100$). For example, the geometry of a small detail in the object would be completely invisible in each of the sub-images. To be able to infer on very high resolution images, we propose here to use a multi-scale approach as introduced in Hardy et al. (2023) and to extend it to the Universal problem. In Hardy et al. (2023), we use an architecture based on the convolution network proposed by Chen et al. (2018). It consists of an encoder and a decoder. Each input image is processed independently by the encoder. Then, the extracted feature maps are synthesized using max pooling to create a single feature map

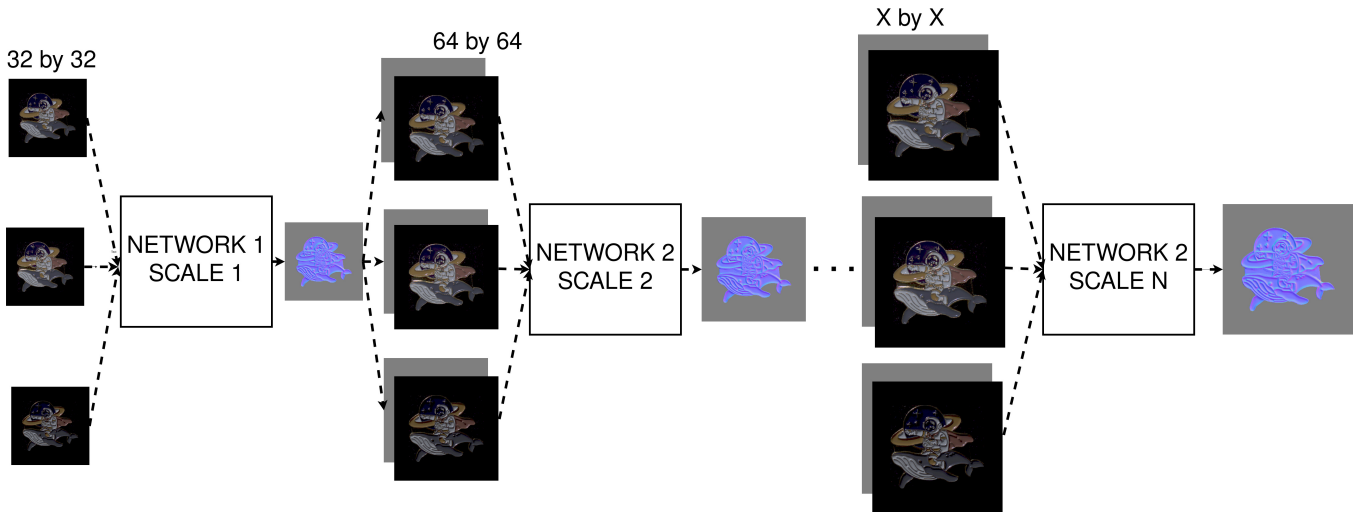


Fig. 3: Global architecture of the proposed multi-scale method. The network can be either our calibrated Transformer, or our Universal Transformer. The first scale is independant from the other scale, then the others scales share parameters.

for all images. The decoder takes this feature map to generate an estimation of the normal map. In the next section, we show how to enhance this architecture in a multi-scale manner, and extend it to the universal setup by relying on Transformers.

3. Proposed Multi-Scale Method

In this section, we first present our multi-scale architecture, and then introduce our new synthetic training dataset.

We propose a multi-scale framework to perform equally well on both low-frequency geometry and high-frequency details. Our multi-scale approach progressively refines the estimated normal map as the spatial scale increases. Our model starts by focusing on the global aspect of the object and then progressively refines details such as holes, cracks, or slight bumps. While a model should be indeed able to handle high-resolution images and different resolutions, a model with a fixed number of convolution layers may still lack sufficient convolutions to effectively synthesize information across an entire arbitrary large image.

To address this, we separate the first scale of our network from the other scales. The first network, used for the first scale, takes as input the downsampled images at the resolution 32×32 pixels and estimates a first normal map at the same resolution. Actually, the first network is used to predict a normal map which can be seen as a initialization prediction for the second network. The latter is used for the remaining scales of our multi-scale approach. It iteratively refines the normal map estimation each time by a factor of 2. Thus, it takes as input an upsampled version of the normal map estimated at the lower resolution as well as images downsampled to the same resolution. A new normal map estimation is predicted at the same resolution as the input images. This process is repeated until we get a normal map at the same resolution of the original images. The key point is that both networks (i.e., for first scale and other scales) have exactly the same architecture, but with different weights. It is necessary to have two independent architectures because the first scale takes as input only the images, while the other scales take as input the concatenation of the images and an estimation of the normal maps. The weights of the second network are identical for every scale.

Our overall architecture is presented in Figure 3.

3.1. Transformers-based Multi-Scale Architecture for Universal PS

Our proposed architecture is a multi-scale encoder-decoder backbone composed of Pyramid Vision Transformer (PVT) blocks [Wang et al. (2021)], Self Attention Blocks (SAB) [Lee et al. (2019)] and Pooling by Multihead Attention (PMA) blocks [Lee et al. (2019)].

Additionally, to test the robustness and performance of our architecture on different PS problems, we propose a variant for solving the calibrated problem. The only difference is the first convolutional embedding layer, which is modified to take either only the images for Universal PS or the images concatenated with the lighting directions for calibrated PS.

Scales Architecture of the Encoder Part — Here we detail the structure of one scale of our proposed method. Each scale is composed of the same encoder and the same decoder. The encoder part combines three modules: the first one extracts the spatial information for each image independently, the second one extracts the lighting information for all images at each pixel location, and the third one pools the information for the skip connections. The decoder part is mainly composed of regression modules.

The spatial extractor module is based on the PVT (Pyramid Vision Transformer) [Wang et al. (2021)]. Indeed, this kind of architecture generates high-resolution features and also features at different scales, allowing us to consider problems at the pixel level. The main advantage is the ability to take in input images of different sizes while keeping moderate computation times. This last point is very important for the photometric

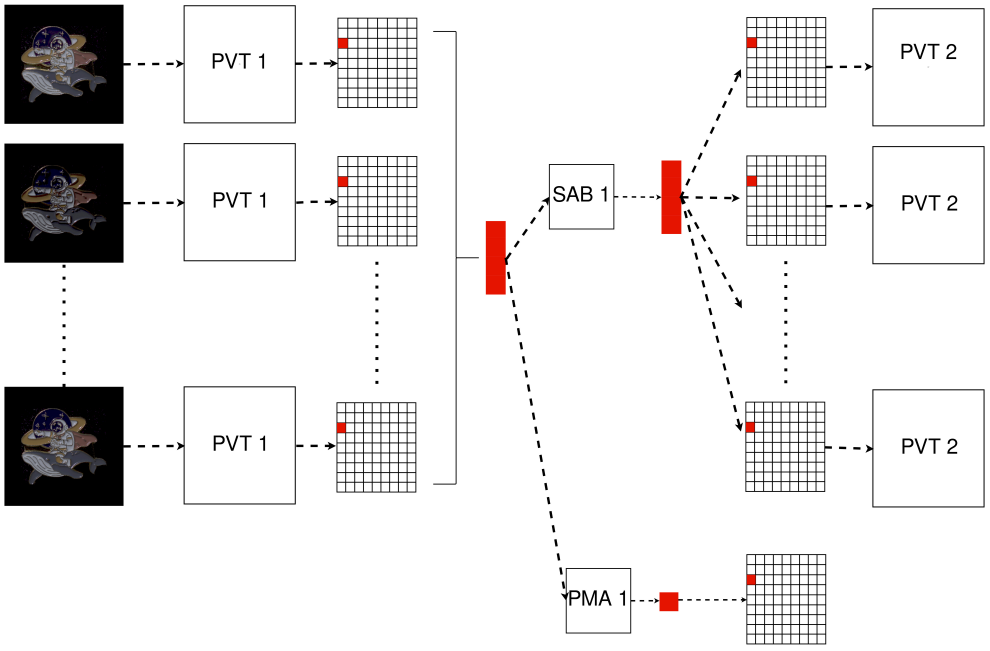


Fig. 4: Detailed architecture of one scale in the encoder part of our proposed method. We process images at pixel level in order to catch all geometric details.

stereo problem, because it is necessary to consider the full size of the images to get a better reconstruction of the normal map.

Then, the lighting extractor module extracts light information at the pixel level. To do so, we use a Self Attention Block (SAB) module. Indeed, at a fixed pixel location, we concatenate the pixel value of each image in order to merge the information at this location. Therefore, we can apply the attention block at each pixel location independently.

Finally, we use a Pooling by Multihead Attention (PMA) module in parallel with the SAB module to aggregate the information given by the PVT block. The aim is to create a feature map and use it for the skip connection in the decoder part.

Figure 4 presents the architecture of the encoder for one scale of our proposed method.

Scales Architecture of the Decoder Part — Once the four encoding blocks are processed, the normal maps are reconstructed with the decoder. For the decoder, we consider three transposed convolutions with skip connections to the PMA map. Indeed, at each step, we concatenate the PMA map obtained in the encoder with the output of the transposed convolution, and so on until we have the desired resolution. The final step consists of a 3×3 convolution to fuse the first PMA map with the output of the last transposed convolution without changing the shape of the feature map, and to create the final normal map.

The decoder architecture is presented in Figure 5.

3.2. Training Dataset

To obtain the best normal map reconstruction possible, a proper dataset needs to be used for the training stage. Most available training datasets are built for photometric stereo in dark environments with parallel light beams [Chen et al. (2018), Ikehata (2018)]. For Universal PS, Ikehata introduced the PS-Wild training dataset [Ikehata (2022c)]. Unfortunately, this dataset has some issues, such as a lack of diversity in geometry, materials, and environments (see Table 1). This appears to be not enough to calibrate a neural network properly to be able to handle all possible materials and geometries.

Training database	samples	shapes	materials	ambient environments
PS-Wild	10 099	410	926	31
Our training database	100 000	11 000	200 000	1 100

Table 1: Comparison beetween our training dataset and PS-Wild. Our training dataset proposes more objects with a larger variety of geometries, shapes and environment than PS-Wild [Ikehata (2022c)].

To solve these issues, we create a new training dataset. To do so, we render 11,000 diverse objects from Scan the world (2023), Sketchfab, and MyMiniFactory using the Blender software [Blender-Foundation]. To complete the lack of smooth surfaces that can exist on these types of objects, we also generate 3,000 distinct objects using the sum of random Gaussian potentials and the Marching Cubes algorithm [Lorensen and Cline (1987)] to extract isosurfaces.

Each time we render a scene, we apply a random material to the object. For materials, we use more than 1,100 'real' materials taken from AmbientCG and around 200,000 materials from Deep-materials [Deschaintre et al. (2018)]. Moreover, we

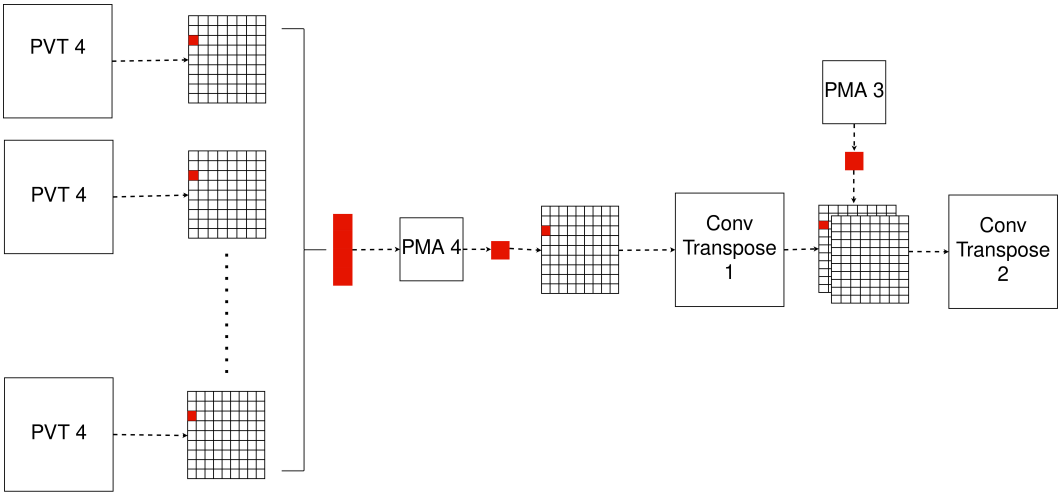


Fig. 5: Architecture of the decoder for one scale of the proposed method. The encoder part is detailed in Figure 4 and the decoder part is composed of transposed convolution.

generate random materials to complete all possible materials. These new materials are created with the BRDF layer of the Cycle render engine [Cycles-developper] in Blender by choosing random values as inputs of this layer.

Finally, for the ambient lighting environment, we use 1,100 360° HDR (High Dynamic Range) images from diverse sources, such as Poly-Haven (2023), AmbientCG and Alexandre Duret-Lutz (2023).

For each sample, we render 50 images with random light distribution over the hemisphere. To ensure that our model can handle different types of lights, we use directional lights and non-parallel lights. Each time the non-parallel light type is chosen for the scene, the size of the bulb and other light parameters are also chosen randomly. On the contrary, the power of the light varies between each image, regardless of the type of lamp chosen. In total, we generate 100,000 samples. A comparison between our training dataset and PS-Wild is shown in Table 1.

Examples of training images generated by our pipeline are given in Figure 6. As we can see, materials, shapes, geometries and environments are diverse.

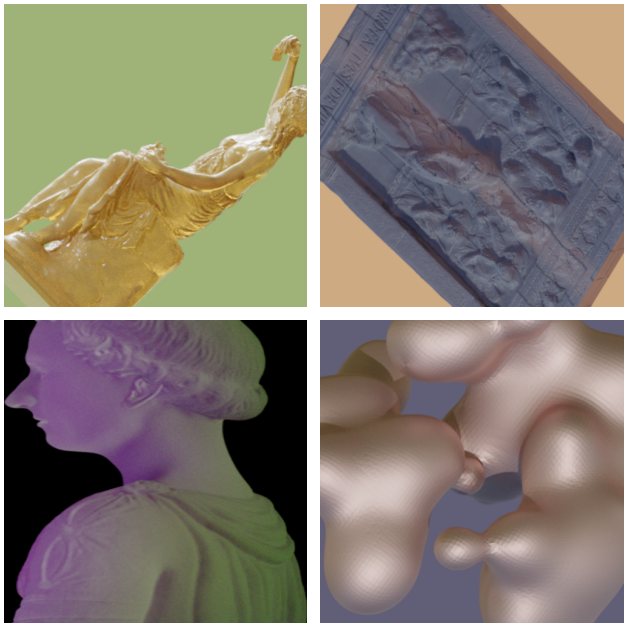


Fig. 6: Examples of training images generated by our pipeline. We can see the variety of geometries, shapes and environments.

3.3. Training Process

To train our multi-scale netwok, we use images with a resolution of 128 by 128 pixels. To reach the resolution of 128 pixels, 3 stages are necessary: 32 by 32 pixels, 64 by 64 pixels, and 128 by 128 pixels. Because of the small resolution of our training images, we are able to give 23 images per view during the training process. A batch size of 2 is enough for the training. Therefore, we are able to use a single A100 to train our method. The Adam optimizer is used with a learning rate of 10^{-4}. The three stages are trained together, and the cosine similarity loss is used. This loss measures the angular difference between the estimated 3D normals and the ground truth normals. It is defined as:

$$l_{normal} = \sum_{ij} \left(1 - N_{ij} \cdot \hat{N}_{ij}\right), \quad (1)$$

where $\hat{N_{ij}}$ is the predicted normal at pixel (i, j), and N_{ij} is the ground truth normal. Everything was implemented with the Pytorch framework.

3.4. Inference on Very High Resolution Image

As mentioned previously, our method can be used for any size of input image. However, performing inference on very high-resolution images is challenging, because even with a batch size of 1, the image may not fit on a single graphics card.

Therefore, to run our network on very high-resolution images, we embed our multi-scale approach in a patch-based heuristic. For images up to 256x256 pixels (i.e., 32 by 32, 64 by 64, 128 by 128, and 256 by 256), we use the full resolution. For larger images, we cut each image and its corresponding predicted normal map into 256x256 patches with an overlap of 64 pixels. We then process each patch independently. Finally, we merge all patches together using a spatial weighted average, with Gaussian weights as defined below:

$$w(x, y) = e^{-\frac{\|(x,y)-(x_c,y_c)\|^2}{2\sigma^2}}, \quad (2)$$

where (x_c, y_c) is the center of the patch and $\sigma = 25$. This value has been chosen empirically.

This method allows us to avoid computing the attention map on the full resolution image. Indeed, computing the attention map can significantly increase the memory requirement in the PVT module when the image size increases.

However, the results in Table 2 show that performance can degrade if the patch size is too small. Therein, we tested our network combined with the proposed patch-based inference on DiLiGenT10^2 [Ren et al. (2022)] at the full image resolution with 30 images per object.

Finally, we choose to take a patch size equal to 256 as it seems to be a reasonable compromise between memory cost and accuracy.

Patch size (px)	Overlap (px)	Memory usage (Go)	MAE (°)
128	32	3.5	15.52
256	64	21	13.19
512	128	130	12.96

Table 2: Memory usage and mean angular error of the proposed patch inference method on the full image resolution of DiLiGent10^2. A patch size of 256 px seem to be a good compromise between performance and memory cost.

4. Experiments

From a quantitative point of view, we compare our approach with all state-of-the-art methods. Indeed, we consider calibrated [Ikehata (2018); Chen et al. (2022); Honzatko et al. (2021); Logothetis et al. (2021); Ju et al. (2022); Lichy et al. (2022); Logothetis et al. (2022); Hardy et al. (2023)], uncalibrated [Chen et al. (2019); Li and Li (2022); Lichy et al. (2022)] and universal methods [Ikehata (2022c, 2023)].

4.1. Description of the testing datasets

We evaluate quantitatively our method on the three publicy available dataset with directional light DiLiGenT [Shi et al. (2016)], DiLiGenT10^2 [Ren et al. (2022)] and DiLiGenT-Pi [Wang et al. (2023)]. We also test the generalization capacity of our method on a dataset with non-parallel light directions named Luces [Mecca et al. (2021)]. Examples of images of each dataset are presented in Figure 7.

• The DiLiGenT dataset [Shi et al. (2016)] is a real-world image dataset containing 96 images from the same viewpoint captured under known light directions and light intensities. It contains 10 objects with ground truth normal maps, obtained by scanning objects with a 3D scanner.

• DiLiGenT10^2 [Ren et al. (2022)] contains 10 objects, each manufactured in 10 different materials. The diversity of materials in this dataset is quite large. Indeed, diffuse, moderately specular, metallic materials with anisotropic reflectance, and translucent materials are all present in this dataset. This diversity offers the possibility to test the performance of our algorithms on challenging materials. The ground truth is also available, but was obtained using 3D digital models and not by a 3D scanner as in the DiLiGenT dataset.

• The last dataset with directional light is DiLiGenT-Pi [Wang et al. (2023)]. This dataset was created to test photometric stereo methods on near-planar surfaces with rich details, such as coins and badges. It contains four groups of materials: metallic, specular, rough, and translucent surfaces. In addition, the dataset contains 30 objects, each with 100 photographs provided. As with the DiLiGenT dataset, the ground truth normals were obtained using a scanner.

• Finally, Luces [Mecca et al. (2021)] is a dataset with non-parallel and near-lighting. It contains 14 objects and 52 images per object. The light directions, intensities, and camera calibration are known. As with the other datasets, the ground truth normal maps are available and were obtained using a scanner.

All of these datasets offer the opportunity to test our method on a wide variety of object shapes, materials, contexts, and so on. However, to complete our evaluation and test the performance of our method on different types of light, environments, and cameras, we also visually test the performance on publicly available datasets where no ground truth is given, such as Skoltech3D [Voynov et al. (2023)], Shape and Material [Lichy et al. (2021)], UNI-PS, [Ikehata (2022c)], and SDM-UniPS [Ikehata (2023)]. The image acquisition setup varies from dataset to dataset. In Skoltech3D [Voynov et al. (2023)], an industrial camera is used and images are captured in the dark with directional light. In UNI-PS [Ikehata (2022c)], an 8-bit smartphone camera is used with near light (within 30 cm of the object) to have spatially varying lighting effects. In SDM-UniPS [Ikehata (2023)] a digital camera is used, and in Shape and Material [Lichy et al. (2021)] an iPhone is used. The variety in the acquisition setup is important for testing the generalization capability of our method to any setup or environment.

4.2. Quantitative comparison

We first evaluate our methods on DiLiGenT in Table 3. We compare the performance of both our Universal and calibrated

(a) DiLiGenT [Shi et al. (2016)] (b) DiLiGenT10^2 [Ren et al. (2022)] (c) DiLiGenT-Pi [Wang et al. (2023)] (d) Luces [Mecca et al. (2021)]

Fig. 7: Examples of images of DiLiGenT [Shi et al. (2016)], DiLiGenT10^2 [Ren et al. (2022)], DiLiGenT-Pi [Wang et al. (2023)] and Luces [Mecca et al. (2021)] datasets.

	type	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	average
PS-FCN [Chen et al. (2022)]	C	2.67	7.72	7.52	4.75	6.72	7.84	12.39	6.17	7.15	10.92	7.39
CNN-PS [Ikehata (2018)]	C	2.2	4.6	7.9	4.1	8.0	7.3	14.0	5.4	6.0	12.6	7.2
OB-Cnn [Honzatko et al. (2021)]	C	2.49	3.59	7.23	4.69	4.89	6.89	12.79	5.10	4.98	11.08	6.37
PX-NET [Logothetis et al. (2021)]	C	2.03	3.58	7.61	4.39	4.69	6.90	13.10	5.08	5.10	10.26	6.28
NormAttention-PSN [Ju et al. (2022)]	C	2.93	5.48	7.12	4.65	5.99	7.49	12.28	5.96	6.42	9.93	6.83
Our previous method, MS-PS [Hardy et al. (2023)]	C	2.05	4.24	7.03	3.9	4.00	7.57	11.01	4.94	5.22	8.47	5.84
SDPS-Net [Chen et al. (2019)]	UC	2.8	6.9	9.0	8.1	8.5	11.9	17.4	8.1	7.5	14.9	9.5
SCPS-NIR [Li and Li (2022)]	UC	1.24	3.82	9.28	4.72	5.53	7.12	14.96	6.73	6.50	10.54	7.05
UNI-PS [Ikehata (2022c)]	UC/Uni	4.9	9.1	19.4	13.0	11.6	24.2	25.2	10.8	9.9	18.8	14.7
SDM-UniPS [Ikehata (2023)]	UC/Uni	**1.5**	3.6	7.5	5.4	4.5	8.5	10.2	4.7	4.1	8.2	5.8
Our (K=30)	C	1.93	2.64	5.88	3.05	3.76	6.40	10.44	3.85	4.32	7.31	4.96
Our (K=96, all images)	UC/Uni	1.92	3.14	6.16	3.60	4.04	**6.35**	**8.84**	**4.08**	4.88	7.09	5.01
Our (K=30)	UC/Uni	1.84	3.14	**6.04**	**3.45**	**3.99**	6.49	8.9	4.12	**4.7**	**7.0**	**4.97**
Our (K=15)	UC/Uni	1.93	**3.05**	6.31	3.97	4.06	7.0	9.27	4.25	4.9	7.41	5.22
Our (K=6)	UC/Uni	2.4	3.7	7.14	4.52	4.7	8.06	12.43	5.32	5.84	9.4	6.35
Our (K=3)	UC/Uni	3.58	4.83	11.46	7.13	6.68	17.8	18.05	8.79	7.75	15.65	10.17

Table 3: Mean angular error (in degrees) on the DiLiGenT benchmark [Shi et al. (2016)]. The type C means calibrated PS, UC is uncalibrated PS and Uni is Universal PS as defined in Ikehata (2022c). The best result is indicated in bold, and the second best one is underlined. The proposed method gives best state-of-the-art results.

Transformer methods with calibrated, uncalibrated, and Universal state-of-the-art methods for PS. Our Universal method clearly outperforms all other methods by at least 16% on all objects. Our calibrated method also achieves state-of-the-art performance compared with calibrated methods. Interestingly though our Universal method reaches results comparable to the calibrated one.

In addition, we test our proposed Universal method with different numbers of images (K=3, 6, 15, 30, and 96). We can see that with only 6 images, our method obtains results that are close to the state-of-the-art using all the available images. Moreover, the results are already the best with only 30 images.

Then, we compare our methods on a more challenging dataset, DiLiGenT10^2 [Ren et al. (2022)], in Table 4. On this dataset, we can see that our Universal method still achieves state-of-the-art results. Our Universal method improves the state-of-the-art results of Universal PS by 13%, from 14.96° to 13.19°. On difficult geometries, like Turbine, the improvement is significant compared to other methods, even compared to calibrated ones. Also we still obtain good results on specular material like AL (aluminium), CU (brass) or STEEL , contrary to other Universal PS methods.

Our transformer calibrated method is the best performer on this dataset (see Table 5). We note that all multi-scale methods get much better results than non-multi-scale methods like CNN-PS [Ikehata (2018)] which is the best performer of the non-multi-scale methods.

The second challenging dataset is DiLiGenT-Pi [Wang et al. (2023)]. Again, our Universal method outperforms all other Universal and uncalibrated methods, see Table 6. Compared to the calibrated methods, our Universal method tends to have slightly lower performance on near-flat objects, but it still

mean: 14.96

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	1.7	1.2	2.5	2.8	2.7	2.4	2.8	5.0	6.9	3.6
GOLF	12.0	6.2	13.0	5.0	12.0	7.1	6.7	6.3	7.5	9.2
SPIKE	12.0	6.8	11.0	6.5	8.7	7.6	8.4	5.7	9.2	13.0
NUT	16.0	5.1	18.0	4.7	8.4	4.8	18.0	13.0	7.5	20.0
SQUARE	23.0	5.4	25.0	5.3	12.0	7.3	19.0	5.5	20.0	32.0
PENTAGON	24.0	8.1	29.0	8.4	13.0	10.0	25.0	26.0	25.0	29.0
HEXAGON	18.0	6.5	20.0	4.8	11.0	7.0	20.0	14.0	19.0	34.0
PROPELLER	28.0	8.3	44.0	6.1	24.0	7.2	22.0	12.0	19.0	28.0
TURBINE	46.0	10.0	51.0	9.4	36.0	11.0	31.0	25.0	25.0	31.0
BUNNY	36.0	8.9	44.0	6.3	19.0	8.1	27.0	7.8	11.0	28.0

(a) SDM-UniPS [Ikehata (2023)] (Universal)

mean: 13.19

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	7.4	6.3	7.5	8.4	9.9	7.2	9.0	9.8	13.0	43.0
GOLF	14.0	7.6	14.0	5.3	11.0	7.6	8.0	8.2	9.6	38.0
SPIKE	9.3	7.1	10.0	6.7	7.4	6.7	11.0	7.7	13.0	29.0
NUT	16.0	6.5	20.0	5.8	8.8	8.1	9.9	10.0	9.1	35.0
SQUARE	18.0	6.0	21.0	7.4	9.1	6.5	6.7	5.5	7.4	35.0
PENTAGON	20.0	8.8	24.0	9.9	12.0	9.7	13.0	13.0	12.0	40.0
HEXAGON	17.0	8.2	18.0	5.5	11.0	6.2	9.4	8.0	7.9	41.0
PROPELLER	15.0	8.1	20.0	6.2	8.4	7.3	16.0	9.2	12.0	21.0
TURBINE	30.0	9.9	33.0	9.6	15.0	10.0	22.0	13.0	17.0	33.0
BUNNY	15.0	8.0	23.0	6.0	7.2	8.1	10.0	7.8	8.2	12.0

(b) Our Universal transformer

Table 4: Mean angular error (in degrees) on the DiLiGenT10^2 benchmark, with the results of SDM-UniPS [Ikehata (2023)] indicated for comparison. The lower the values, the better the results are. Our Universal method gives best state-of-the-art results.

achieves competitive results. We note that the average is not necessarily the best metric to compare the performance of calibrated and uncalibrated methods. Indeed, for some objects all uncalibrated or Universal state-of-the-art methods predict an inverted normal map compared to the ground truth (for example, see Figure 9). This is likely because uncalibrated methods are unable to determine the direction of incoming light and tend to assume that it is coming from the opposite direction to the actual direction. As shown in Figure 9a, it is difficult to tell if the light is coming from above or below. Both possibilities are equally plausible, but would result in opposite normal maps.

Our methods are way more robust to this problem than other uncalibrated and Universal methods, as we only have 2 objects inverted compared to 11 for SDM-UniPS [Ikehata (2023)] and 8 for SDPS-NET [Chen et al. (2019)]. Indeed, as shown in Figure 10, our uncalibrated methods is able to predict correctly the normal map contrary to SDM-UniPS [Ikehata (2023)] and Uni-PS [Ikehata (2022c)].

mean: 15.78

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	5.1	6.4	4.2	4.5	6.9	7.3	16.0	14.0	16.0	19.0
GOLF	14.0	8.0	12.0	6.8	14.0	9.4	12.0	9.2	13.0	22.0
SPIKE	11.0	9.4	11.0	11.0	12.0	9.5	14.0	8.3	16.0	28.0
NUT	20.0	8.8	19.0	6.9	17.0	8.0	16.0	13.0	14.0	22.0
SQUARE	21.0	8.1	22.0	6.7	19.0	8.1	13.0	4.9	7.9	18.0
PENTAGON	26.0	9.5	26.0	9.8	22.0	9.6	15.0	13.0	15.0	23.0
HEXAGON	18.0	7.5	19.0	7.2	17.0	28.0	18.0	10.0	17.0	21.0
PROPELLER	28.0	12.0	35.0	8.4	23.0	11.0	16.0	9.6	9.8	17.0
TURBINE	54.0	20.0	51.0	16.0	39.0	21.0	25.0	22.0	21.0	32.0
BUNNY	24.0	11.0	27.0	7.8	21.0	9.1	12.0	7.7	12.0	14.0

(a) CNN-PS [Ikehata (2018)] (calibrated)

mean: 11.33

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	9.3	3.4	8.7	5.2	8.4	4.3	8.5	12.0	14.0	8.6
GOLF	10.0	7.3	9.8	5.8	10.0	6.87	7.9	7.7	9.8	12.0
SPIKE	12.0	8.8	9.9	6.3	8.5	7.9	12.0	7.6	12.0	17.0
NUT	14.0	8.9	15.0	5.8	10.0	5.8	9.2	7.6	8.2	16.0
SQUARE	18.0	11.0	17.0	8.2	14.0	5.5	12.0	7.2	7.9	11.0
PENTAGON	18.0	8.4	17.0	8.0	17.0	9.4	11.0	9.4	13.0	20.0
HEXAGON	16.0	7.5	15.0	6.1	13.0	7.1	11.0	8.1	11.0	20.0
PROPELLER	13.0	8.9	11.0	7.9	16.0	9.7	11.0	8.4	8.2	19.0
TURBINE	21.0	12.0	24.0	11.0	18.0	15.0	23.0	16.0	18.0	22.0
BUNNY	17.0	8.2	16.0	6.5	12.0	8.3	8.6	7.3	8.0	18.0

(b) Our previous method, MS-PS [Hardy et al. (2023)] (calibrated)

mean: 11.01

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	4.5	3.3	5.0	3.6	5.8	4.1	3.6	7.8	8.8	6.4
GOLF	13.0	6.4	14.0	5.1	11.0	6.8	7.0	6.4	8.1	9.3
SPIKE	10.0	7.3	11.0	7.5	9.1	8.3	8.5	8.3	9.0	11.0
NUT	11.0	5.0	19.0	4.5	8.1	5.0	6.6	6.8	6.4	24.0
SQUARE	18.0	8.5	23.0	7.7	13.0	7.1	7.9	5.0	7.6	19.0
PENTAGON	15.0	8.3	22.0	8.9	13.0	8.4	11.0	9.5	9.5	21.0
HEXAGON	16.0	5.8	20.0	5.9	12.0	6.4	7.1	5.2	6.7	20.0
PROPELLER	16.0	9.2	32.0	8.2	9.6	6.9	15.0	7.8	10.0	17.0
TURBINE	30.0	9.4	30.0	10.0	19.0	9.8	22.0	15.0	16.0	23.0
BUNNY	12.0	8.0	29.0	6.0	8.0	8.3	9.4	8.5	8.4	13.0

(c) Our calibrated transformer

Table 5: Mean angular error (in degrees) on the DiLiGenT10^2 benchmark, with the results of CNN-PS [Ikehata (2018)] and our previous multi-scale CNN [Hardy et al. (2023)] indicated for comparison. The lower the values, the better the results are. Our calibrated Transformer method gives best state-of-the-art results.

	Type	Astro Lung	Bagua-R Ocean	Bagua-T Panda-R	Bear Panda-T	Bird Para	Cloud-R Queen	Cloud-T Rhino	Crab Sail	Fish Ship	Flower Sun	Lion-R TV	Lion-T Taichi	Lions Tree	Lotus-R Wave	Lotus-T Whale	average
NormAttention-PSN [Ju et al. (2022)]	C	7.2 7.8	12.0 5.8	16.5 13.9	7.4 16.6	6.9 4.2	13.4 **4.9**	17.3 5.1	**4.4** 5.2	4.4 **4.9**	4.6 **5.6**	16.4 **7.6**	21.0 9.7	**4.4** 9.6	10.8 6.1	13.7 8.7	9.2
PS-FCN [Chen et al. (2018)]	C	7.2 9.7	13.0 5.8	16.8 14.8	7.4 17.2	7.2 4.7	14.3 **4.7**	17.8 5.3	5.3 **5.1**	**4.6** 6.1	4.6 6.7	18.4 8.0	21.2 10.2	4.5 10.6	11.8 6.8	13.6 12.2	9.85
CNN-PS [Ikehata (2018)]	C	6.0 5.7	12.2 **4.6**	16.4 14.2	7.4 16.6	6.8 **3.9**	14.6 5.4	17.2 **4.9**	4.5 5.2	4.2 4.9	4.7 5.8	15.8 8.3	20.3 7.8	4.7 11.3	10.9 5.3	13.5 11.6	9.16
Our previous method, MS-PS [Hardy et al. (2023)]	C	**5.96** 7.51	11.32 4.97	15.1 14.75	6.9 14.72	7.69 4.09	13.28 6.37	14.74 5.18	4.58 5.26	4.68 **5.14**	5.43 6.46	14.37 8.63	15.71 9.91	5.5 8.22	11.92 5.29	12.8 7.09	8.78
SDPS-Net [Chen et al. (2019)]	UC	37.7 40.2	22.5 31.4	28.9 21.8	30.7 23.7	17.6 19.8	27.4 16.5	27.5 24.9	20.5 16.7	23.6 19.0	12.8 31.5	20.8 26.9	23.6 34.1	19.6 41.1	21.7 39.1	26.5 29.8	25.93
SDM-UniPS [Ikehata (2023)]	UC/Uni	37.8 46.6	14.6 34.6	17.1 17.1	23.8 17.6	26.5 23.2	17.1 10.6	19.2 17.0	25.4 10.5	24.5 22.0	15.2 26.2	15.9 36.6	16.2 47.2	9.2 34.4	11.8 34.9	13.6 33.8	23.34
Our (k=30)	C	6.03 **5.41**	**9.57** 5.44	11.75 **12.98**	**6.72** **11.39**	**6.55** 4.73	12.61 5.69	**11.01** 5.22	5.75 6.66	**4.11** 6.25	4.85 5.9	13.12 10.24	11.43 **7.26**	5.37 **6.08**	**10.17** 5.48	**8.09** **6.71**	**7.75**
Our (k=100, all images)	UC/Uni	7.58 42.1	10.19 6.35	**11.12** 13.5	12.49 11.9	8.14 7.2	**12.45** 7.43	11.63 6.69	6.0 7.2	8.32 5.35	5.88 6.54	12.66 10.39	11.24 8.54	6.63 47.27	11.29 6.11	10.38 7.84	11.35
Our (k=30)	UC/Uni	7.14 41.98	10.43 5.91	11.69 13.28	14.09 12.22	7.35 7.13	13.08 9.54	11.92 6.68	5.32 6.62	5.96 5.65	5.14 6.05	**12.73** 11.5	**11.2** 8.95	6.16 47.15	11.51 5.93	10.39 8.77	11.38
Our (k=15)	UC/Uni	10.93 43.73	10.46 10.46	13.44 13.81	12.16 12.65	8.21 7.34	12.71 7.68	14.04 6.83	8.23 8.2	8.76 5.99	8.02 8.05	14.19 11.37	12.2 10.45	7.31 48.9	11.79 7.49	11.19 9.49	12.54

Table 6: Mean angular error (in degrees) on the DiLiGenT-Pi benchmark [Wang et al. (2023)]. Best results are in bold, and the second best ones are underlined. The type C means calibrated PS, UC is uncalibrated PS and Uni is Universal PS as defined in Ikehata (2022c). The best result is indicated in bold, and the second best one is underlined. The proposed method gives best state-of-the-art results.

		Ball	Bell	Bowl	Buddha	Bunny	Cup	Die	Hippo	House	Jar	Owl	Queen	Squirrel	Tool	average
Fast-PS (v1) [Lichy et al. (2022)]	C	8.55	**6.20**	7.0	12.69	8.63	17.28	**5.16**	8.01	29.00	**5.32**	12.32	12.90	13.00	12.33	11.32
L22 [Logothetis et al. (2022)]	C	8.84	7.51	**5.95**	**11.59**	**7.06**	15.35	5.19	**5.60**	**22.97**	6.19	**8.89**	**9.97**	11.77	**11.64**	**9.90**
Fast-PS (v2) [Lichy et al. (2022)]	UC	**6.59**	7.17	10.17	14.50	11.75	18.98	8.63	10.64	31.00	9.14	15.92	18.39	15.97	18.61	14.11
UNI-PS [Ikehata (2022c)]	UC/Uni	11.012	24.12	23.84	27.90	23.51	28.64	16.24	21.41	35.93	14.53	32.87	28.36	25.36	19.03	23.77
SDM-UniPS [Ikehata (2023)]	UC/Uni	13.30	12.76	8.44	18.58	8.53	19.67	7.25	8.86	26.07	8.30	12.67	15.97	16.01	12.54	13.50
Our (K=52, all images)	UC/Uni	10.20	10.52	6.98	12.83	9.60	13.68	6.19	8.33	25.29	6.30	11.47	12.45	**11.36**	11.79	11.21
Our (K=30)	UC/Uni	10.29	10.51	6.79	12.57	9.6	**13.35**	6.27	8.44	25.46	6.10	11.38	15.97	11.37	12.22	11.10
Our (K=15)	UC/Uni	10.47	10.8	7.91	13.14	9.90	13.96	6.52	8.54	25.30	6.49	11.82	12.49	11.64	11.89	11.50
Our (K=6)	UC/Uni	10.94	11.40	9.38	13.75	11.029	15.38	7.80	9.41	26.68	7.37	12.62	12.85	12.79	12.47	12.42
Our (K=3)	UC/Uni	10.93	15.95	12.07	16.78	14.53	16.09	9.09	11.06	31.61	10.49	15.73	14.99	15.67	15.69	15.05

Table 7: Mean angular error (in degrees) on the Luces benchmark [Mecca et al. (2021)]. Best results are in bold, and the second best ones are underlined. The type C means calibrated PS, UC is uncalibrated PS and Uni is Universal PS as defined in Ikehata (2022c). The best result is indicated in bold, and the second best one is underlined. The proposed method gives best state-of-the-art results.

		Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	average
Without mask	SDM-UniPS [Ikehata (2023)]	4.42	4.21	8.54	5.59	7.24	10.37	14.92	5.44	6.72	12.97	8.04
	Our Universal	11.46	4.64	7.46	4.11	7.80	7.14	10.34	5.27	5.59	7.93	7.17
With mask	SDM-UniPS [Ikehata (2023)]	1.5	3.6	7.5	5.4	4.5	8.5	10.2	4.7	4.1	8.2	5.8
	Our Universal	1.84	3.14	6.04	3.45	3.99	6.49	8.9	4.12	4.7	7.0	4.97

Table 8: Mean angular error (in degrees) on the DiLiGenT benchmark [Shi et al. (2016)] without masking the background before processing. For comparison the results with background is also shown.

		Ball	Bell	Bowl	Buddha	Bunny	Cup	Die	Hippo	House	Jar	Owl	Queen	Squirrel	Tool	average
Without mask	SDM-UniPS [Ikehata (2023)]	10.45	14.27	10.94	21.29	11.91	10.69	7.56	9.34	27.47	7.33	13.69	16.23	17.16	15.37	13.84
	Our Universal	14.6	13.95	8.29	12.5	8.81	9.19	8.54	9.78	25.52	9.61	12.49	12.26	12.44	16.95	12.49
With mask	SDM-UniPS [Ikehata (2023)]	13.30	12.76	8.44	18.58	8.53	19.67	7.25	8.86	26.07	8.30	12.67	15.97	16.01	12.54	13.50
	Our Universal	10.20	10.52	6.98	12.83	9.60	13.68	6.19	8.33	25.29	6.30	11.47	12.45	11.36	11.79	11.21

Table 9: Mean angular error (in degrees) on the Luces benchmark [Mecca et al. (2021)] without masking the background before processing. For comparison the results with background is also show.

In this dataset, our calibrated Transformer gives also very good results. Overall, our calibrated Transformer method gives a significant improvement of 12% compare to the second best. And again, all our multi-scale architectures obtain the best results in their categories.

Finally, our Universal method is able to also manage non-parallel light beam as showned in Table 7. We obtain the best performance over other Universal PS methods and uncalibrated methods especially built for non parallel scenario. Compared to the best calibrated results our method is very close in term of performances. Our method reachs comparable result with Lichy et al. (2022), yet these methods use an explicit lighting model whose parameters must be calibrated (location, intensities, dispertion,...).

Inference with no mask — One advantage of SDM-UniPS of Ikehata (2023) compared to the other PS methods is its ability to solve the PS problem without using any object mask. This is a novel feature for deep learning-based methods, as common deep learning PS methods require to mask the background to work properly. To test this type of inference only the quantitative datasets DiLiGenT [Shi et al. (2016)] and Luces [Mecca et al. (2021)] are suitable for use, as the backgrounds in the other datasets are completely dark or cropped.
We infer the normal map without masking the background and then compute the normal error only on the object part. This technique allows us to test the impact of masking the background for our Universal method. The performance decreases without masking the background (see Table 8 and Table 9), but our method still maintains really good results and outperforms SDM-UniPS [Ikehata (2023)], the only one so far able to manage inference with no mask. In Figure 8, we show an example on two objects: the Reading object from DiLiGenT [Shi et al. (2016)] and the Alligator from SDM-UniPS [Ikehata

(2023)]. We can see that our Universal method not only generates a proper normal map for the desired object, but also reconstructs the background correctly, which is not necessarily the case for SDM-UniPS [Ikehata (2023)]. For example, the carpet is reconstructed much better with our Universal method.

SDM-UniPS [Ikehata (2023)] Our

Fig. 8: Comparison on the Alligator object of SDM [Ikehata (2023)] and the Reading objet of DiLiGenT [Shi et al. (2016)] without masking background. We can see that considering the background does not degrade reconstruction of normals. We reconstruct more accurate details than SDM [Ikehata (2023)].

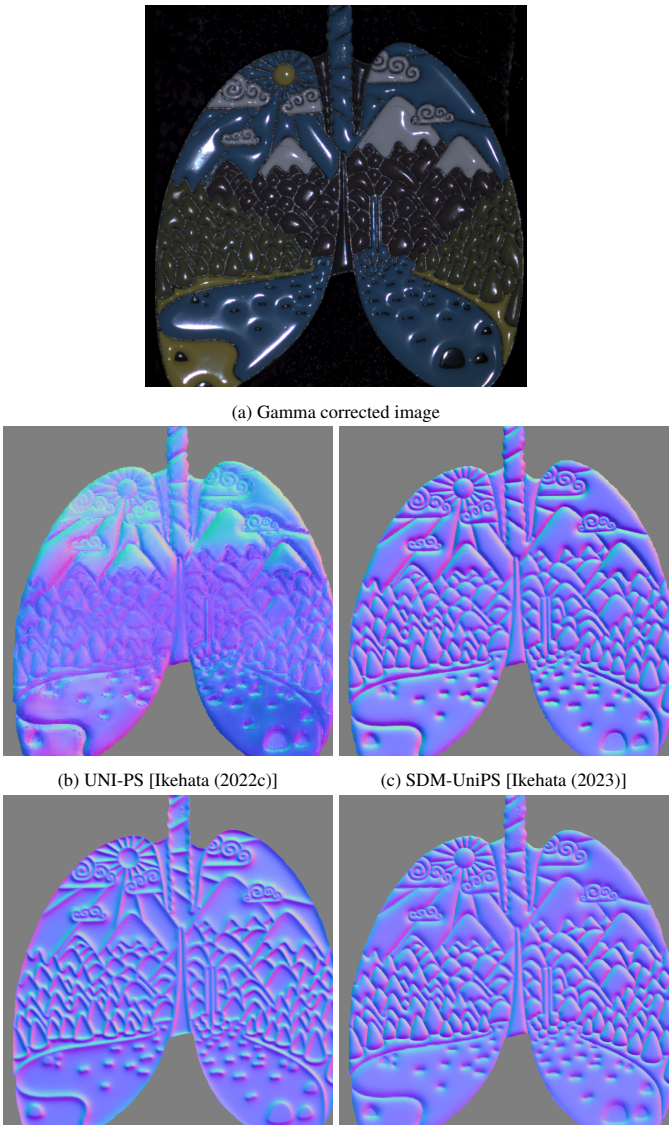
4.3. Qualitative evaluation

Next, to test the robustness of our Universal method in the most diverse contexts and environments, we use several available qualitative datasets. We compare our Universal method only to SDM-UniPS [Ikehata (2023)] as all other methods are not Universal. Indeed, Uni-PS [Ikehata (2022c)] is a Universal PS method, but its performance is below SDM-UniPS on all quantitative datasets, so it is not of interest to compare to it. Note that in this section, we only focus on inference with masked backgrounds.

Overall, the results seem good for both methods, but our Universal method outperforms SDM-UniPS on surface details (see Figures 11 and 12). In the Owl object in Figure 11, the results seem similar, but when zooming on the talons, artifacts appear on the prediction of SDM-UniPS [Ikehata (2023)] which is not the case with our Universal method. Finally, with our multi-scale method, the results remain good regardless of the resolution. For example, in Figure 13 and 14, the images resolution are really high and our Universal method obtains excellent results. Our method performs way better than SDM-UniPS [Ikehata (2023)] especially on nearly 'flat' objects like the ceramic (Figure 13).

4.4. Limitations

The main limitation of Universal/Uncalibrated methods is the normal map reconstruction on translucent material like acrylic. None of the state-of-the-art methods give accurate reconstruction. Indeed, as the material is translucent, it is very difficult to know from which side the light is coming. For example, in some objects like acrylic balls, the light passes through the ball. So it is actually really hard to determine if the light source is located on the left or the right of the ball. Another example is shown in Figure 9a. Without any prior knowledge of the object shape, it is difficult to find out precisely where the light is coming from. This greatly impacts methods for uncalibrated PS, as the two opposite incoming light directions would lead to the perfectly opposite normals. So, our Universal method can be improved on this type of material.

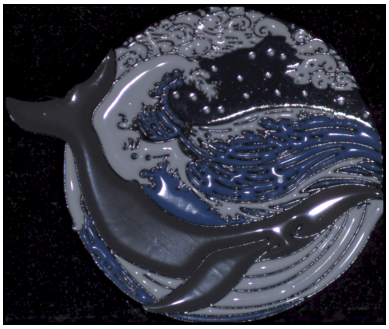


(a) Gamma corrected image

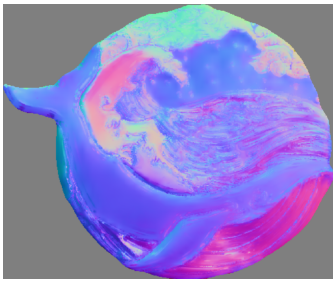
(b) UNI-PS [Ikehata (2022c)] (c) SDM-UniPS [Ikehata (2023)]

(d) Our calibrated Transformer (e) Our Universal Transformer

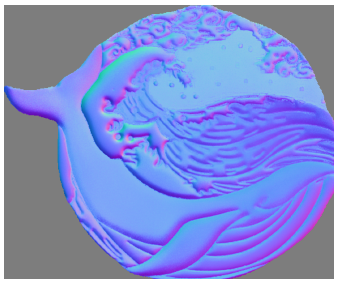
Fig. 9: Normal prediction of our methods, SDM-UniPS [Ikehata (2023)] and Uni-PS [Ikehata (2022c)] on the Lung object of Wang et al. (2023). We can see that this material is challenging for uncalibrated and Universal approaches because of the light reflection. Indeed, normal maps are inverted.



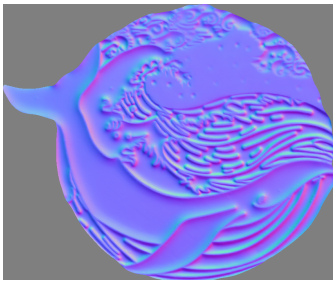
(a) Gamma corrected image



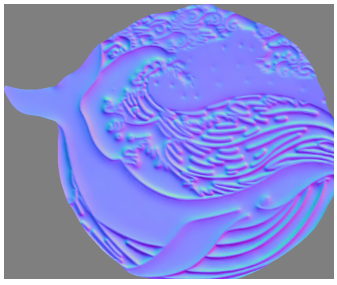
(b) UNI-PS [Ikehata (2022c)]



(c) SDM-UniPS [Ikehata (2023)]



(d) Our calibrated Transformer



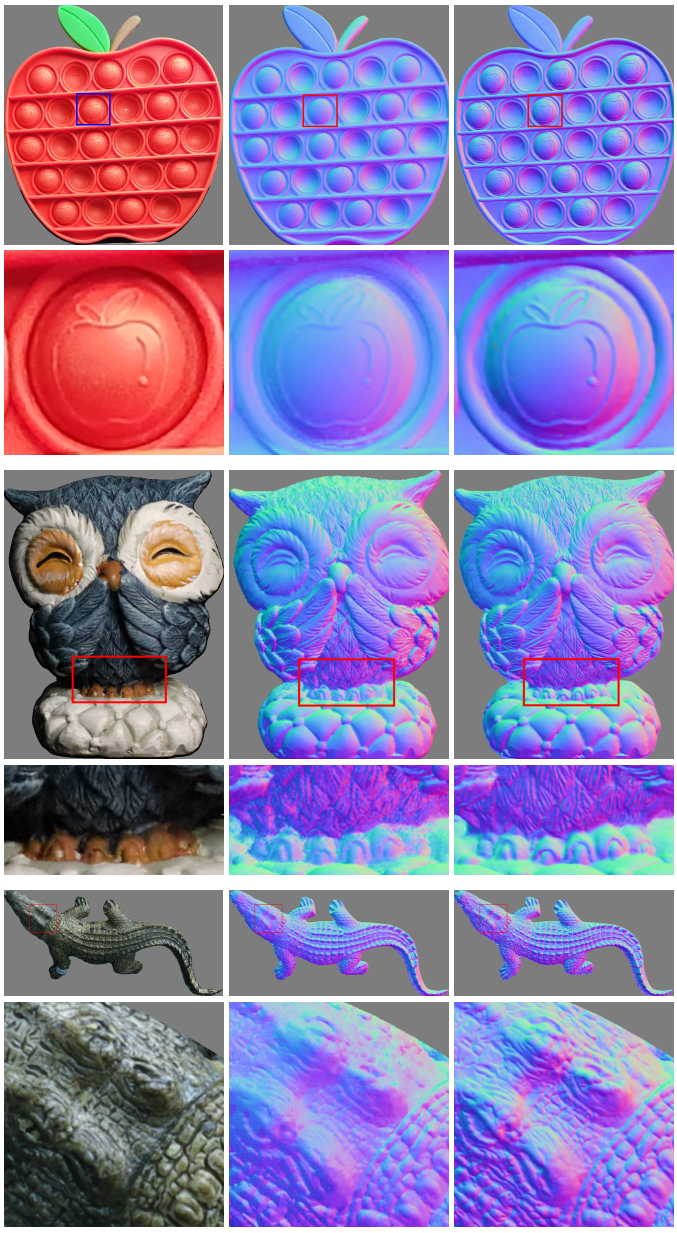
(e) Our Universal Transformer

Fig. 10: Normal prediction of our methods, SDM-UniPS [Ikehata (2023)] and Uni-PS [Ikehata (2022c)] on the Whale object of Wang et al. (2023). Our Universal method gives the correct normal orientation. The other uncalibrated and Universal approaches fail, inverting the normals orientations.

5. Conclusion

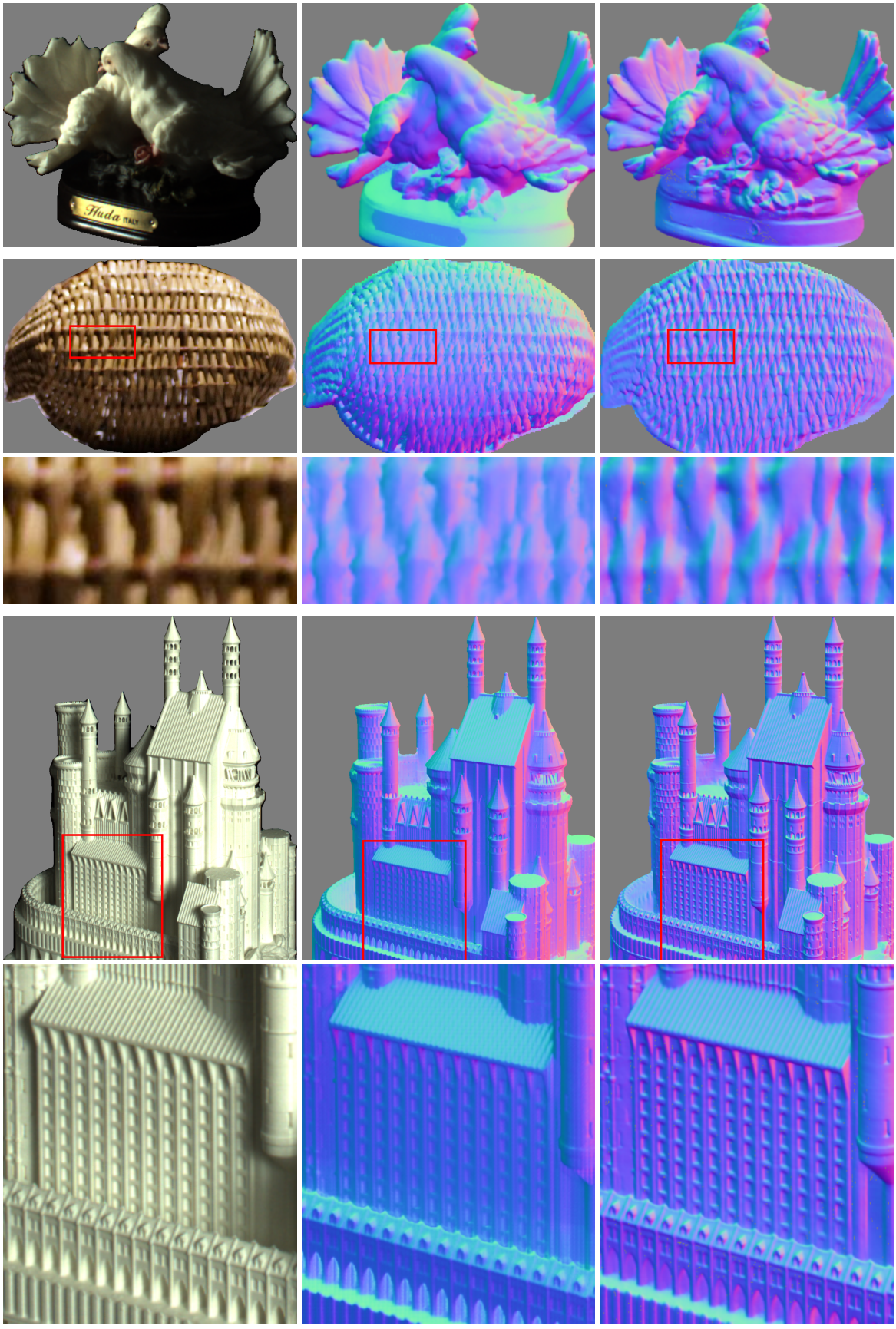
To conclude, we propose a new multi-scale approach based on Transformers with encoder and decoder for each scale. Our method gives excellent results over a large panel of benchmark datasets with a large diversity of acquisition setup and environments which show its robustness. Our method also shows its capacity to manage very high resolution image to get the smallest details of the geometry and to keep very high normal reconstruction performance.

Acknowledgment. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD010613775 made by GENCI.



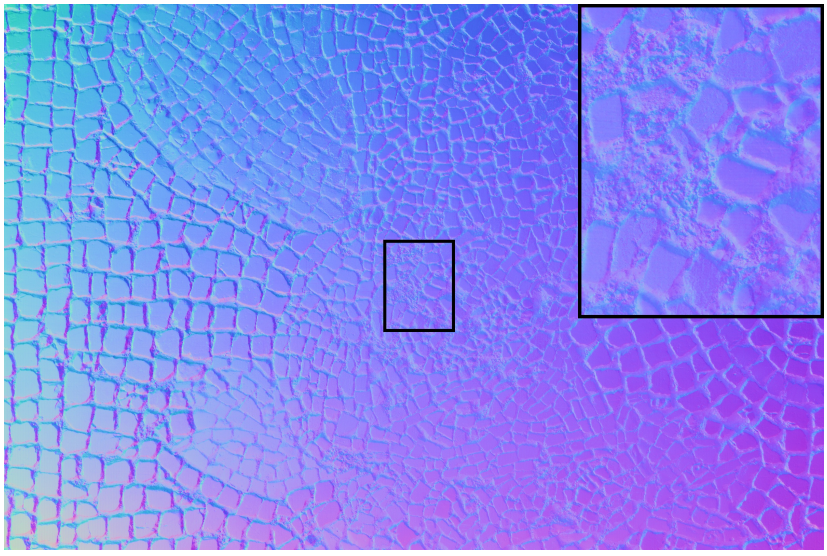
SDM-UniPS [Ikehata (2023)] SDM-UniPS [Ikehata (2023)] Our Universal Transformer

Fig. 11: Comparison on objects without ground truth from Ikehata (2022c, 2023). The first column is the RGB images, the second one is the SDM method [Ikehata (2022c)] and the last one is our method. Then, for all object, we present the full image and a zoom part. For all objects, we reconstruct more accurate details than SDM [Ikehata (2022c)].

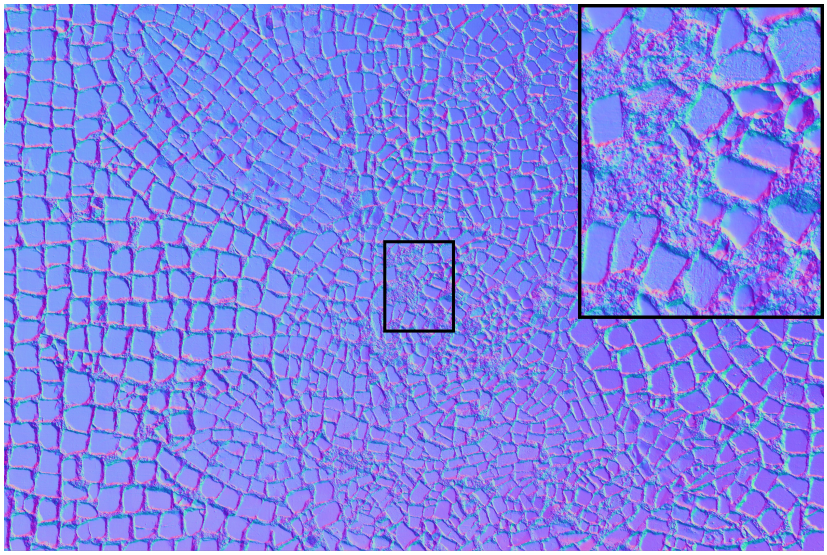


SDM-UniPS [Ikehata (2023)] SDM-UniPS [Ikehata (2023)] Our Universal Transformer

Fig. 12: Comparison on objects without ground truth from Voynov et al. (2023); Lichy et al. (2021). The first column is the RGB images, the second one is the SDM method [Ikehata (2022c)] and the last one is our method. Then, for all object, we present the full image and a zoom part on the next line. For all objects, we reconstruct more accurate details than SDM [Ikehata (2022c)].

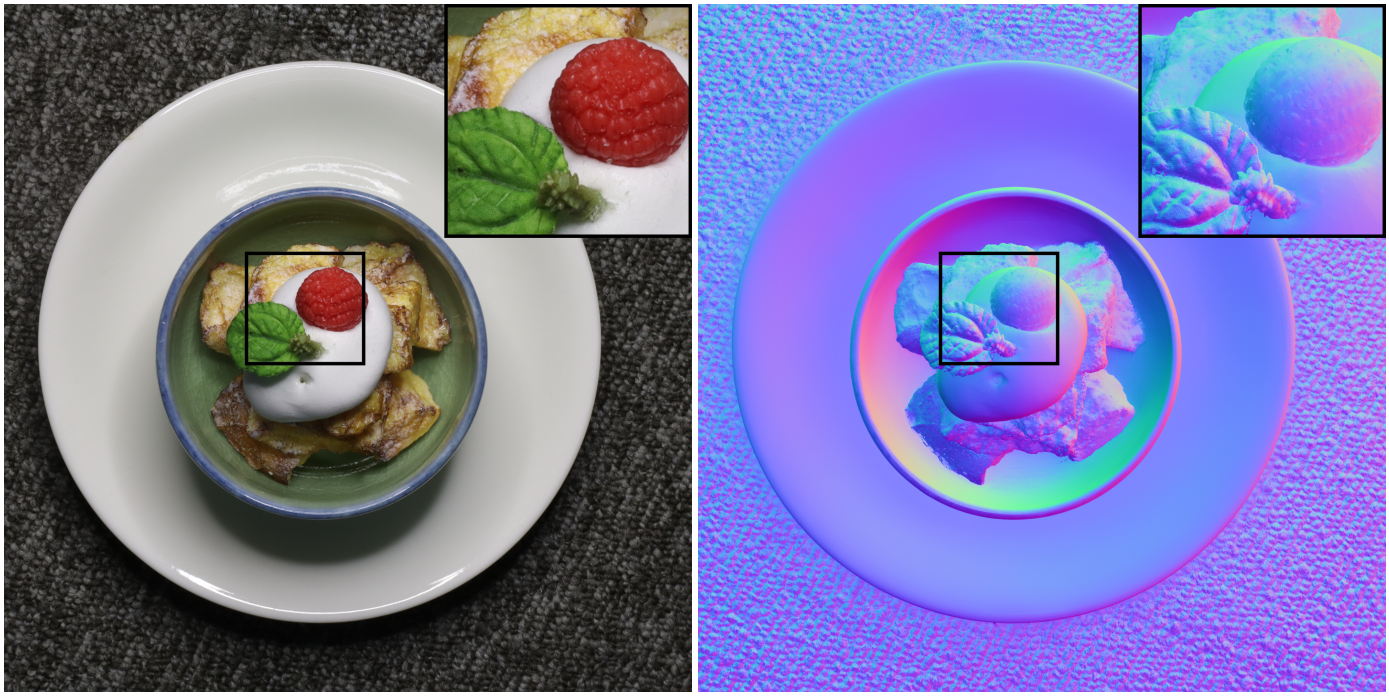


SDM-UniPS [Ikehata (2023)]

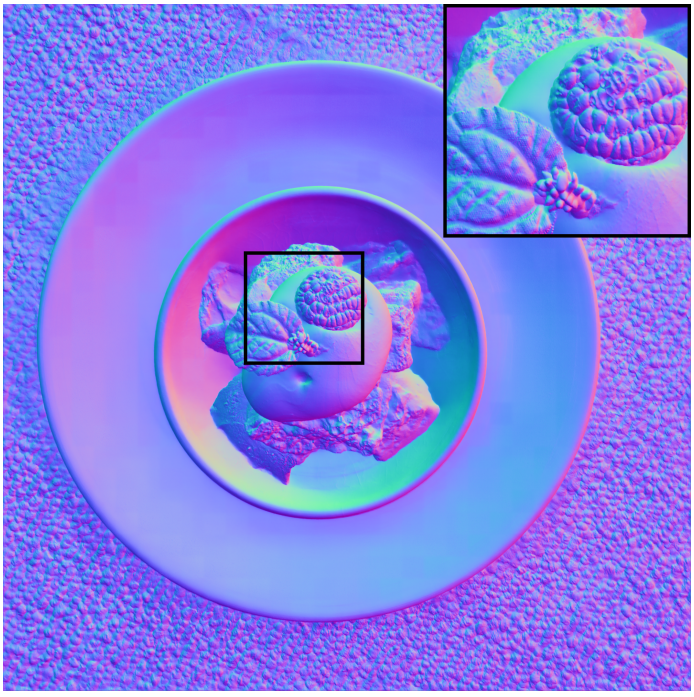


Our

Fig. 13: Visual comparison beetween SDM-UniPS [Ikehata (2023)] method and our Universal method on the Seasons mosaic object. The image resolution is 5 500 by 8 200 pixels. We can see that our method can manage very high images and outperforms SDM-UniPS [Ikehata (2023)] in terms of normal map reconstruction. PS image credits: A. Laurent (INRT, UMR 5055 IRIT)/MAN.



SDM-UniPS [Ikehata (2023)]



Our

Fig. 14: Visual comparison beetween SDM-UniPS [Ikehata (2023)] method and our Universal method on the Sweet object of Ikehata (2023). The image resolution is 4 000 by 4 000 pixels. We can see that our method can manage very high resolution images and outperforms SDM-UniPS [Ikehata (2023)] in terms of normal map reconstruction.

References

Alexandre Duret-Lutz, 2023. URL: https://www.flickr.com/people/gadl/.

AmbientCG, . https://ambientcg.com/.

Blender-Foundation, . Blender - a 3D modelling and rendering package. Stichting Blender Foundation, Amsterdam. URL: http://www.blender.org.

Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.K., 2022. Deep photometric stereo for non-lambertian surfaces. IEEE Transactions on Pattern Analysis amp; Machine Intelligence 44, 129–142.

Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.Y.K.K., 2019. Self-calibrating deep photometric stereo networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8731–8739.

Chen, G., Han, K., Wong, K.Y.K., 2018. Ps-fcn: A flexible learning framework for photometric stereo, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV), Cham. pp. 3–19.

Cycles-developper, . Cycles is a physically based production renderer developed by the Blender project.

Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A., 2018. Single-image svbrdf capture with a rendering-aware deep network. ACM Transactions on Graphics (SIGGRAPH) 37.

Hardy, C., Quéau, Y., Tschumperlé, D., 2023. MS-PS: A Multi-Scale Network for Photometric Stereo With a New Comprehensive Training Dataset, in: International Conference on Computer Graphics, Visualization and Computer Vision (WSCG), pp. 194–203.

Honzatko, D., Turetken, E., Fua, P., Dunbar, L., 2021. Leveraging spatial and photometric context for calibrated non-lambertian photometric stereo, in: Proceedings of the International Conference on 3D Vision (3DV), pp. 394–402.

Ikehata, S., 2018. Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV).

Ikehata, S., 2022a. Does physical interpretability of observation map improve photometric stereo networks?, in: 2022 IEEE International Conference on Image Processing (ICIP), pp. 291–295.

Ikehata, S., 2022b. Ps-transformer: Learning sparse photometric stereo network using self-attention mechanism, in: British Machine Vision Conference (BMCV).

Ikehata, S., 2022c. Universal photometric stereo network using global lighting contexts. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 12581–12590.

Ikehata, S., 2023. Scalable, detailed and mask-free universal photometric stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13198–13207.

Ju, Y., Dong, J., Chen, S., 2021. Recovering surface normal and arbitrary images: A dual regression network for photometric stereo. IEEE Transactions on Image Processing (TIP) , 3676–3690.

Ju, Y., Lam, K.M., Chen, Y., Qi, L., Dong, J., 2020. Pay attention to devils: A photometric stereo network for better details, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI), pp. 694–700.

Ju, Y., Shi, B., Jian, M., Qi, L., Dong, J., Lam, K.M., 2022. Normattention-psn: A high-frequency region enhanced photometric stereo network with normalized attention. International Journal of Computer Vision (IJCV) 130, 3014–3034.

Kaya, B., Kumar, S., Oliveira, C., Ferrari, V., G., V., 2021. Uncalibrated neural inverse rendering for photometric stereo of general surfaces, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3804–3814.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W., 2019. Set transformer: A framework for attention-based permutation-invariant neural networks, in: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the International Conference on Machine Learning (ICML), pp. 3744–3753.

Li, J., Li, H., 2022. Self-calibrating photometric stereo by neural inverse rendering, in: Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV), pp. 166–183.

Li, J., Robles-Kelly, A., You, S., Matsushita, Y., 2019. Learning to minify photometric stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7560–7568.

Li, Z., Zheng, Q., Shi, B., Pan, G., Jiang, X., 2023. Dani-net: Uncalibrated photometric stereo by differentiable shadow handling, anisotropic reflectance modeling, and neural inverse rendering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8381–8391.

Lichy, D., Sengupta, S., Jacobs, D.W., 2022. Fast light-weight near-field photometric stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12602–12611.

Lichy, D., Wu, J., Sengupta, S., Jacobs, D.W., 2021. Shape and material capture at home, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6119–6129.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 11966–11976.

Logothetis, F., Budvytis, I., Mecca, R., Cipolla, R., 2021. PX-net: Simple and efficient pixel-wise training of photometric stereo networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12737–12746.

Logothetis, F., Mecca, R., Budvytis, I., Cipolla, R., 2022. A cnn based approach for the point-light photometric stereo problem. International Journal of Computer Vision (IJCV) 131, 101–120.

Lorensen, W.E., Cline, H.E., 1987. Marching cubes: A high resolution 3d surface construction algorithm, in: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, Association for Computing Machinery. p. 163–169.

Mecca, R., Logothetis, F., Budvytis, I., Cipolla, R., 2021. Luces: A dataset for near-field point light source photometric stereo, in: British Machine Vision Conference (BMVC).

MyMiniFactory, . https://www.myminifactory.com/fr.

Poly-Haven, 2023. Poly Haven is a curated public asset library for visual effects artists and game designers, providing useful high quality 3D assets in an easily obtainable manner. URL: https://polyhaven.com.

Ren, J., Wang, F., Zhang, J., Zheng, Q., Ren, M., Shi, B., 2022. Diligent102: A photometric stereo benchmark dataset with controlled shape and material variation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12571–12580.

Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y., 2017. Deep photometric stereo network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 501–509.

Scan the world, 2023. URL: https://www.myminifactory.com/scantheworld/.

Shi, B., Wu, Z., Mo, Z., Duan, D., Yeung, S., Tan, P., 2016. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3707–3716.

Sketchfab, . https://sketchfab.com.

Voynov, O., Bobrovskikh, G., Karpyshev, P., Galochkin, S., Ardelean, A.T., Bozhenko, A., Karmanova, E., Kopanev, P., Labutin-Rymsho, Y., Rakhimov, R., Safin, A., Serpiva, V., Artemov, A., Burnaev, E., Tsetserukou, D., Zorin, D., 2023. Multi-sensor large-scale dataset for multi-view 3d reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21392–21403.

Wang, F., Ren, J., Guo, H., Ren, M., Shi, B., 2023. Diligent-pi: Photometric stereo for planar surfaces with rich details - benchmark dataset and beyond, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12571–12580.

Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp. 548–558.

Wang, X., Jian, Z., Ren, M., 2020. Non-lambertian photometric stereo network based on inverse reflectance model with collocated light. IEEE Transactions on Image Processing (TIP) 29, 6032–6042.

Woodham, R.J., 1980. Photometric Method For Determining Surface Orientation From Multiple Images. Optical Engineering , 513–531.

Yao, Z., Li, K., Fu, Y., Hu, H., Shi, B., 2020. Gps-net: Graph-based photometric stereo network, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 10306–10316.

Zheng, Q., Jia, Y., Shi, B., Jiang, X., Duan, L., Kot, A., 2019. Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8548–8557.