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

Clément Hardy, Yvain Quéau, David Tschumperlé

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 (calibrated or uncalibrated) 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:

- managing non-Lambertian materials, like specular ones;
- handling complex illumination, including ambient.

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. To this end, we propose a multi-scale architecture combined with an encoder-decoder Transformer 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 Fig. 1. In this example, our algorithm takes 11 images of size 6000×8000 as input, and it recovers all details of the scene, whose appearance is mostly diffuse.

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

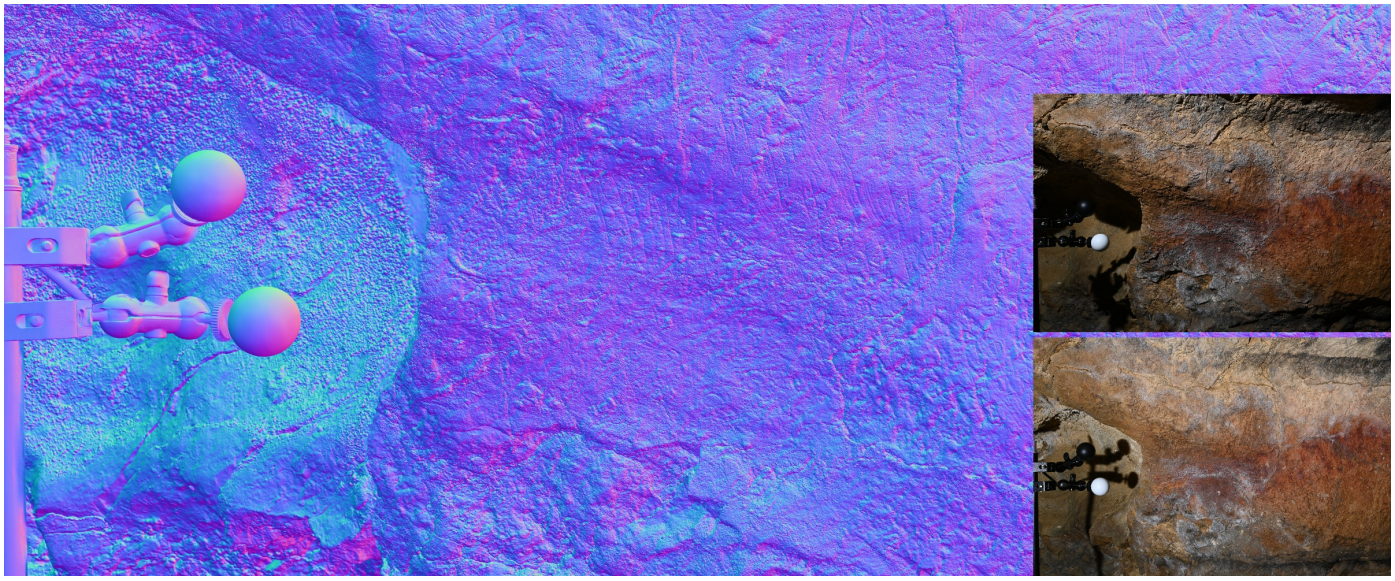


Fig. 1: Reconstruction of the Marsoulas cave, from images of size 6000×8000 pixels. Our method both accomodates the very high image resolution and preserves the fine details. PS image credits: A. Laurent 2023 (INPT, UMR 5505 IRIT), C. Fritz and G. Tosello team (CREAP-E.Cartailhac), MSHS-T (UAR 3414).

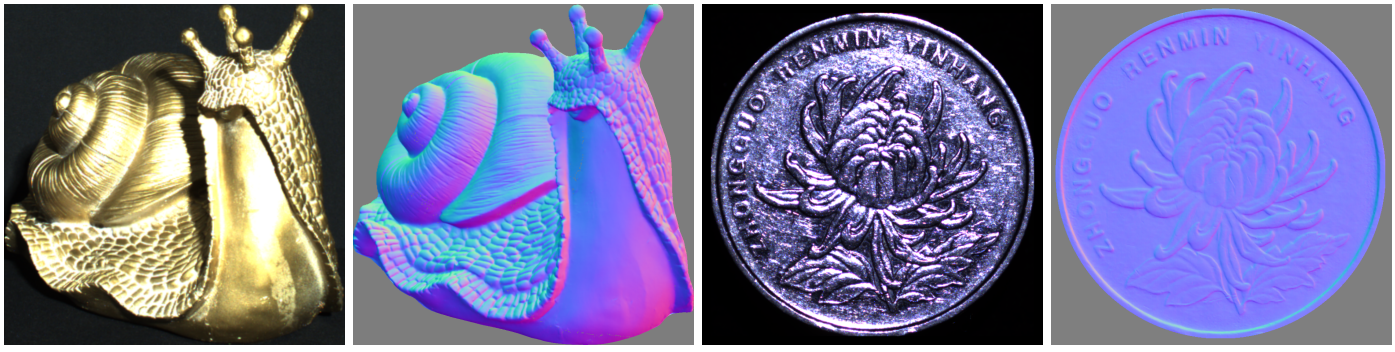


Fig. 2: Reconstruction of two challenging metallic objects (snail from Voynov et al. (2023) and coin from Wang et al. (2023)), with our proposed method.

Our method is also able to manage more difficult materials such as materials, as illustrated in Fig. 2. Besides, it also achieves state-of-the-art results in any environment or lighting conditions, including directional or non-parallel lighting beams, as will be demonstrated on real benchmarking datasets. This last feature constitutes our main improvement over our previous method [Hardy et al. (2023)] – which was designed for calibrated PS in a controlled dark environment. To achieve this, two major changes were made:

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

Let us first review the deep learning methods for PS, according to their illumination requirements.

Calibrated PS. Santo et al. (2017) first proposed a calibrated approach based entirely on a fully connected network. However, light directions must be identical between training and inference, which makes it impractical. Indeed, the neural network architecture should allow an arbitrary number of images, to avoid training a different neural network for each possible number. Two main alternatives have thus been proposed:

- aggregate information from different images using a pooling layer, 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);
- project all the per-pixel observations onto a fixed-size space, as in the work of Ikehata (2018, 2022a), Li et al. (2019), Logothetis et al. (2021) and Zheng et al. (2019).

Per-pixel and all-pixel methods were also unified within a graph-based approach [Yao et al. (2020)]. This is particularly effective under sparse lighting distributions, where Transformers also perform particularly well [Ikehata (2021)].

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, 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)]. 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)]. Such methods optimize an image reconstruction loss (between the reconstructed images and the input images) to get the normals, reflectance and illumination. However, all these uncalibrated methods generalize poorly to natural light/ambient light, because designing a physics-based model for this type of illumination remains difficult. In the traditional (non-deep) context, some attempts towards uncalibrated PS under natural illumination were made, resorting to equivalent directional lighting [Mo et al. (2018)] or spherical harmonics [Haefner et al. (2019)], yet Universal PS methods based on Transformers were recently shown to provide much better results.

Universal PS. Recently, Ikehata introduced the notion of Universal PS with the UniPS [Ikehata (2022b)] 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. They are based on an encoder-decoder model, where the encoder extracts a global lighting context from a fixed 'canonical' resolution image - resizing the images if needed to fit this resolution. 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.

In practice, 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. Combining such downsampling with pixel by pixel inference, very high resolution images can be handled. However, some information is lost during downsampling, and pixel-by-pixel inference lacks spatial information. 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.

Inference on very high resolution images. While UniPS [Ikehata (2022b)] tends to infer blurry and inaccurate normal maps with a lack of detail, especially with high resolution images, SDM-UniPS [Ikehata (2023)] yield much sharper results and scales better. However, it remains difficult to scale to very high resolutions without losing accuracy. 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. An alternative approach consists in resorting to a multi-scale approach, as the one we previously introduced for the calibrated scenario in [Hardy et al. (2023)]. Therein, an architecture based on the convolution network proposed by Chen et al. (2018) was considered. It consists of an encoder and a decoder, where each input image is processed independently by the encoder and the extracted feature maps are then synthesized using max pooling to create a single feature map for all images. The decoder takes this feature map to generate an estimation of the normal map. In the next section, we will extend this multi-scale approach to the Universal setup, by relying on Transformers rather than on CNNs.

Training dataset. In addition to a potential drop in accuracy when resolution increases, the performances of UniPS [Ikehata (2022b)] also tend to decrease with the complexity of materials. Ikehata (2023) explained that this is mostly due to a lack of diversity in the shape and appearance of the objects in the training dataset, and so introduced a new dataset for training UniPS. We had made similar observations in Hardy et al. (2023), showing that the more diverse and representative the training dataset is, the better the results are. Therefore, the present paper also introduces a new dataset which is way more diverse and complete, for training Universal PS methods. Combining this new dataset with the the proposed multi-scale network architecture, which is able to extract both local and global information and thus to much better cope with complex materials[Hardy et al. (2023)], yields state-of-the-art reconstruction results for Universal PS. The next section provides further details on these two contributions.

3. Proposed multi-scale method

In order to be able to infer the smallest details on the object surface, a model should be able to handle both arbitrary high resolution images. Therefore, when using, e.g., a CNN model with a fixed number of convolution layers, this number may lack sufficient convolutions to effectively synthesize information across an entire arbitrary large image. To solve this problematic, we proposed in [Hardy et al. (2023)] a multi-scale framework that performs equally well on both low-frequency geometry and high-frequency details, and can process any size of images. This multi-scale approach progressively refines the estimated normal map as the spatial scale increases. It starts by focusing on the global aspect of the object and then progressively refines details such as holes, cracks, or slight bumps as shown in Fig. 4. Let us start by recalling the global framework of this multi-scale architecture, before specifying it for the Universal case by resorting to Transformers.

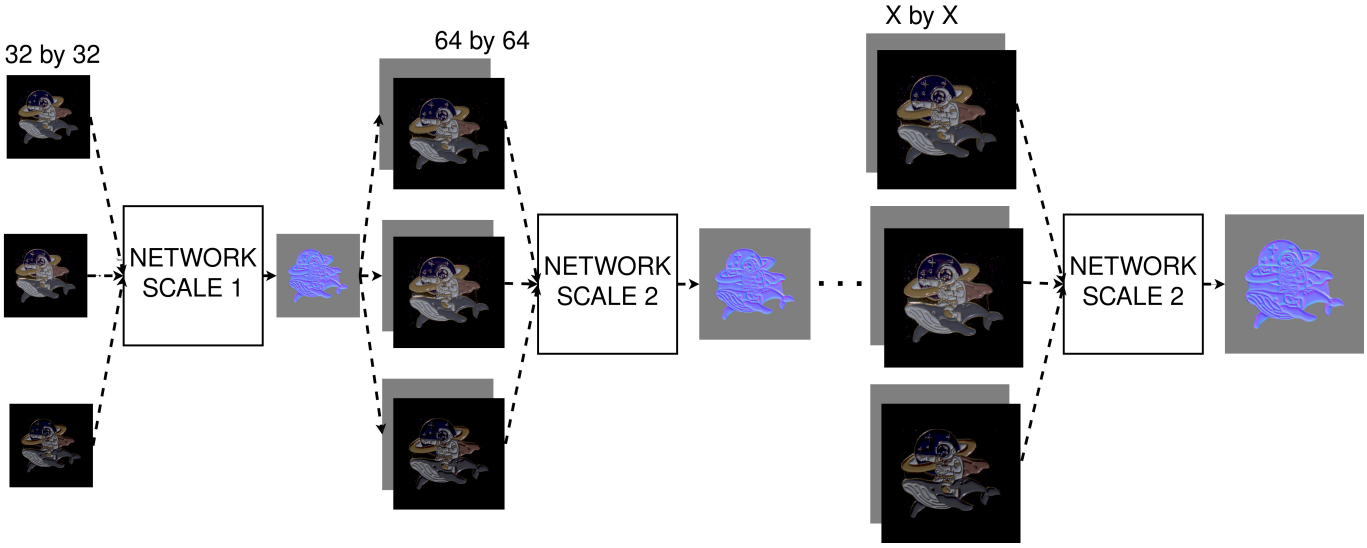


Fig. 3: Global architecture of the proposed multi-scale method. The network at each scale level is an encoder-decoder Transformer detailed in Fig. 5. Let us emphasize that in this architecture, the first scale is independent from the others, which all share the same parameters.

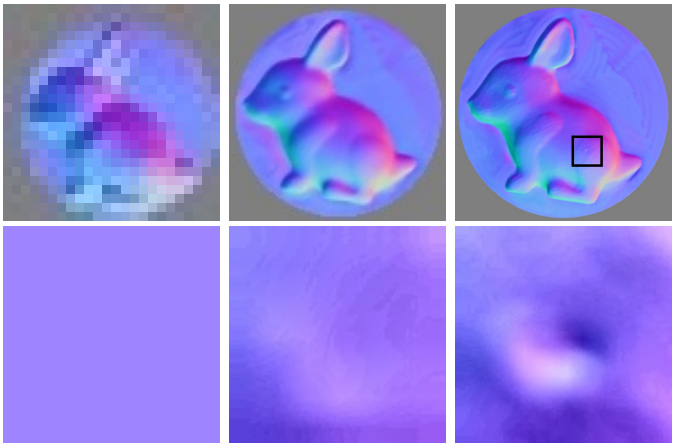


Fig. 4: Detail recovery along with scale level. Contrast is optimized in the zoom area (second line) for visual purpose.

3.1. Global multi-scale architecture

Our overall architecture is presented in Fig. 3. Therein, the first scale is separated from the others. A first network takes as input the downsampled images at the resolution 32×32 pixels and estimates a first normal map at the same resolution. This serves as an initialization prediction for a second network, used for the remaining scales. It iteratively refines the normal map estimation each time by a factor of 2, until the original resolution is reached. 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, and refines the normal map. Both networks (i.e., for first scale and other scales) have exactly the same architecture, but with different weights. It is indeed necessary to have two independent architectures because the first network takes as input only the images, while the other also considers the normals. Nevertheless, the weights of the second network are identical for every scale.

3.2. Transformers-based backbone

Each scale of our architecture involves a network which is essentially an encoder-decoder 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)]. Fig. 5 presents the overall architecture of this network. Depending on which scale is considered, the input to the network is either the images alone, or the images concatenated with the normal maps upsampled from the previous scale. In contrast, our previous work [Hardy et al. (2023)] resorted to CNNs for the backbone, yet this was limited to the calibrated setup and we empirically found out that CNNs generalize poorly to the Universal setup.

To be able to compare Transformers against CNNs, we also designed a variant of the Transformers-based backbone, adapted to the calibrated problem. The only difference is the first convolutional embedding layer of the network, which is modified to take either only the images for Universal PS, or the images concatenated with the lighting directions for calibrated PS. In all cases, each network is composed of the same encoder and decoder architectures, which are detailed hereafter.

Scales architecture of the encoder part. 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 ultimately pools the information for the skip connections.

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

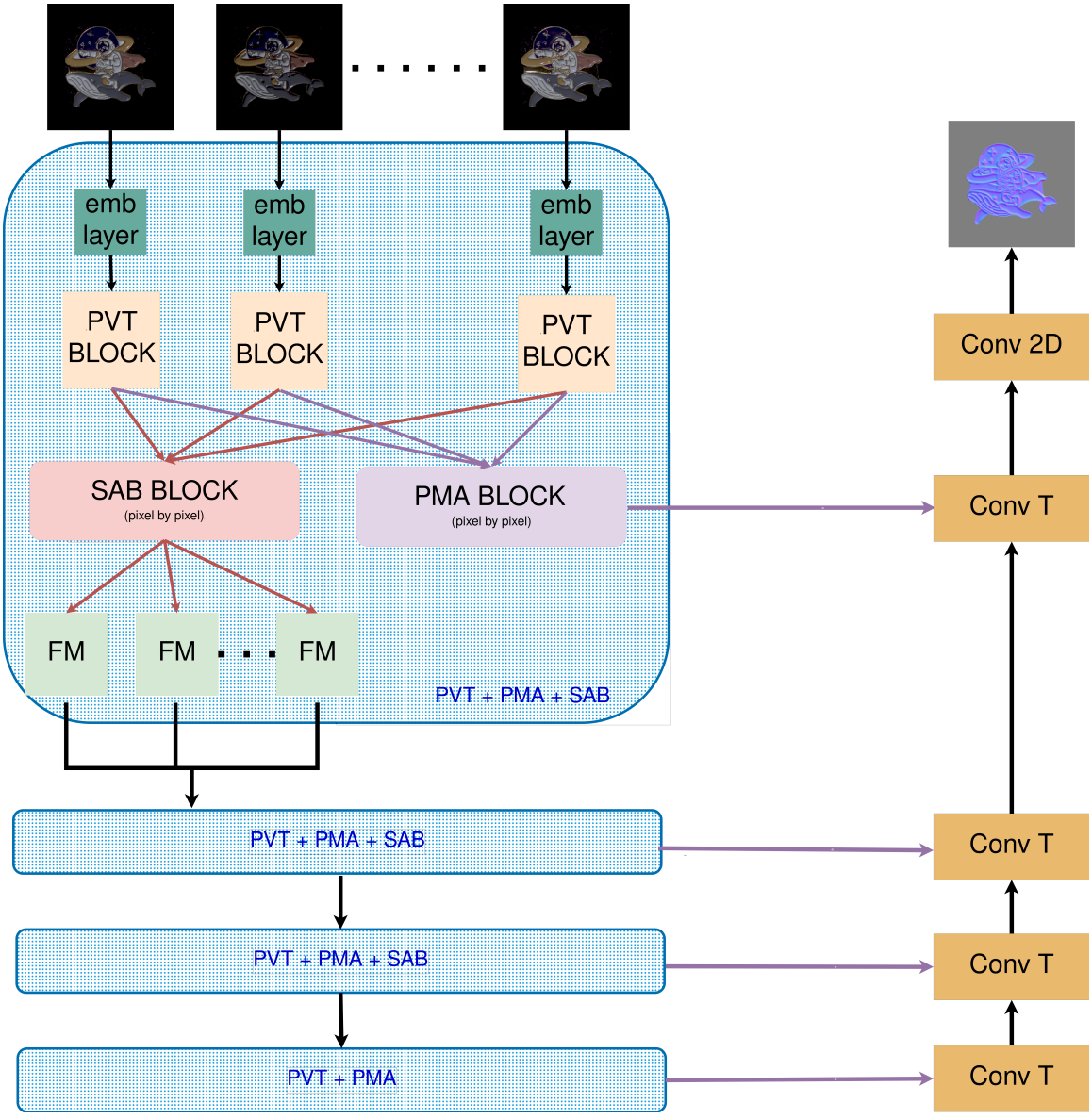


Fig. 5: Detailed architecture of one network scale. The input can be either images alone (for the first network scale), or images concatenated with the upsampled previous normal map (for the other scales). We process images at pixel level in order to catch all geometric details. SAB mean Self Attention Block [Lee et al. (2019)], PMA mean Pooling by Multihead Attention [Lee et al. (2019)], and PVT mean Pyramid Vision Transformer [Wang et al. (2021)]

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. This way, the decoder can utilize information from different extraction levels, helping it retain maximum detail. Additionally, using a pooling layer allows for a variable number of input images as shown in Chen et al. (2018).

Scales architecture of the decoder part. Once the four encoding blocks are processed, the normal maps are reconstructed with the decoder, which is mainly composed of regression modules. We considered 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.

Network size. In total, our proposed multi-scale architecture has 79, 151, 174 parameters, each of the first and second network having around 39.5 millions parameters. Table 3 details the number of parameters for each block.

Block name	input size	output size	nb params
emb layer 1	3/6	64	1,920/3,648
PVT 1	64	64	944,064
SAB 1	64	64	87,936
PMA 1	64	64	166,592
emb layer 2	64	128	74,112
PVT 2	128	128	1,862,144
SAB 2	128	128	347,904
PMA 2	128	128	529,280
emb layer 3	128	256	295,680
PVT 3	256	256	6,372,864
SAB 3	255	256	1,383,936
PMA 3	256	256	1,844,480
emb layer 4	512	512	1,181,184
PVT 4	512	512	9,513,984
PMA 4	512	512	6,834,176
Conv T 1	512	512	4,194,816
Conv T 2	768	256	3,145,984
Conv T 3	384	128	786,560
Conv	192	3	5,187

Table 1: Input, output size and number of parameters for each block of the network. Note that the first embedding layer has two possible input size: 3 for the first network and 6 for the second one.

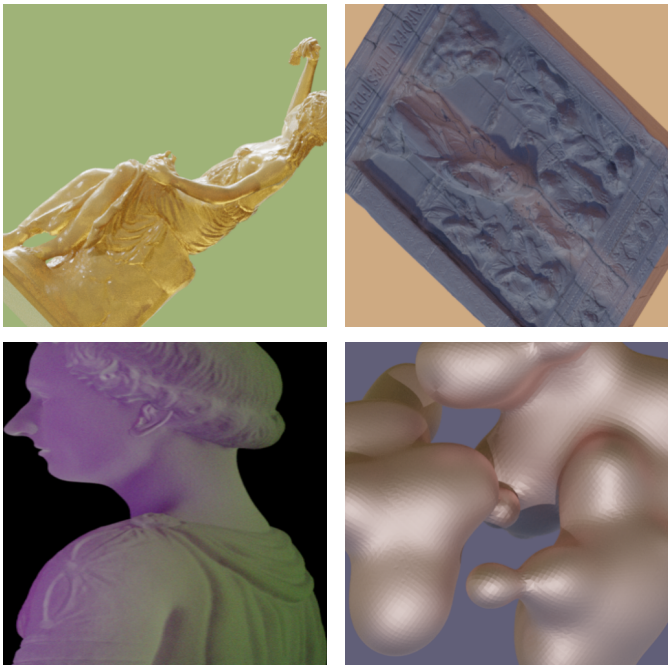


Fig. 6: Examples of training images generated by our pipeline.

3.3. 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 (2022b)]. Unfortunately, this dataset has some issues, such as a lack of diversity in geometry, materials, and environments (see Table 2). 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 2: 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 (2022b)].

To solve these issues, we create a new training dataset. To do so, we render 14,000 diverse objects from the "Scan the world" and "Sketchfab" websites, using the Blender software. 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 the "Ambientcg" website, as well as around 200,000 materials from Deep-materials [Deschaintre et al. (2018)]. Moreover, we generate random materials to complete all possible materials. These random materials were created with the BRDF layer of the Cycle rendering engine in Blender, choosing random values as inputs of this layer.

For each combination of shape and material, 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 addition to these active illuminations, an ambient lighting environment was introduced, by considering 1,100 360° HDR (High Dynamic Range) images from diverse sources, such as "Polyhaven", "Ambientcg" and Alexandre Duret-Lutz's Flickr webpage. In total, we generate 100,000 samples. But as we create a generation pipeline of synthetic images, the total number of samples could have been much larger, as we can create as many samples as needed for our model. A comparison between our training dataset and PS-Wild is shown in Table 2. Examples of training images generated by our pipeline are also given in Fig. 6, highlighting the diversity of materials, shapes, geometries and environments.

3.4. 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 (80G0) to train our method, and it takes roughly three days to train it. 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, which measures the angular difference between the estimated 3D normals and the ground truth normals. Everything was implemented with the Pytorch framework.

3.5. Inference on very high resolution images

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, using Gaussian weights with a standard deviation of 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 3 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. In view of these results, we chose a patch size of 256 pixels as it offers a reasonable compromise between memory usage and accuracy.Indeed, increasing the patch size from 256 to 512 pixels improves performance on DiLiGenT10^2 by only 1.7%, while requiring six times more memory. Furthermore, the visual differences between these two patch sizes are not perceptible, even on a highly specular material that requires the full context of the image to understand the light beams' paths (see Fig. 7).

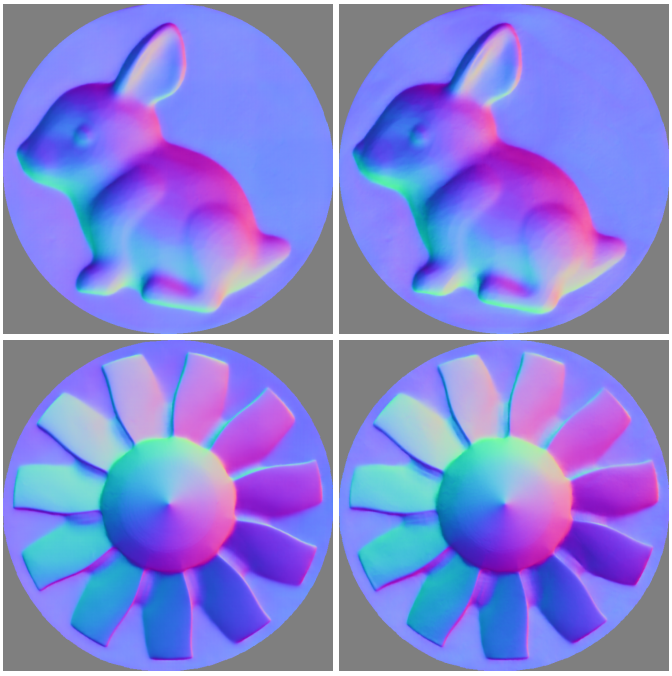
Patch size (px)	Overlap (pixels)	Memory usage (Go)	MAE (°)
128	32	3.5	15.52
256	64	21	13.19
512	128	130	12.96

Table 3: Memory usage and mean angular error of the proposed patch inference method on the full image resolution of DiLiGent10^2 with 15 images as inputs. A patch size of 256 pixels seems to be a good compromise between performance and memory cost.

The memory footprint and processing time also depend on the number of input images. Table 4 illustrates this dependency for a patch size of 256 by 256 pixels on 1000 by 1000 pixel images, indicating the GPU memory footprint and time required for inference for various numbers of input images.

Number of images	Memory usage (Go)	Time computation (seconds)
3	4.27	66
6	7.5	120
9	12.3	173
15	21	280
32	34	560

Table 4: Memory usage and time computation on the full image resolution of DiLiGent10^2 with a patch size of 256 pixels.



Patch size: 256 pixels Patch size: 512 pixels

Fig. 7: Comparison of our universal method between a patch size of 256 pixels and a patch size of 512 pixels. Visually, the difference between the two is slight; however, inference with a patch size of 512 uses six times more GPU memory. Both the bunny and the turbine are in a very specular material, aluminium for the bunny and brass for the turbine.

4. Experiments

Our approach was compared against all state-of-the-art methods, including calibrated [Ikehata (2018); Chen et al. (2022); Honzátko et al. (2021); Logothetis et al. (2021); Ju et al. (2022); Lichy et al. (2022); Logothetis et al. (2023); Hardy et al. (2023)], uncalibrated [Chen et al. (2019); Li and Li (2022); Lichy et al. (2022)] and universal ones [Ikehata (2022b, 2023)].

4.1. Description of the testing datasets

This comparison was carried out 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 Fig. 8.

• DiLiGenT [Shi et al. (2016)] (Table 5) 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)] (Tables 6 and 7) 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. The ground truth is also available, but it was obtained using 3D digital models and not by a 3D scanner as in the DiLiGenT dataset.

(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. 8: Examples of images from the four benchmarking 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 [Honzátko 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
UniPS [Ikehata (2022b)]	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 5: 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 (2022b). The best result is indicated in bold, and the second best one is underlined. The proposed method gives best state-of-the-art results.

• The last dataset with directional light is DiLiGenT-Pi [Wang et al. (2023)] (Table 8). 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)] (Table 9) is a dataset with non-parallel and near-lighting. It contains 14 objects and 52 images per object, with known light locations and intensities. 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)], UniPS, [Ikehata (2022b)], 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 UniPS [Ikehata (2022b)], 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 5. We compare the performance of both our Universal and calibrated Transformer methods with calibrated, uncalibrated, and Universal state-of-the-art methods for PS. Our Universal method outperforms all other methods by at least 16% on all objects. Interestingly, it reaches results comparable to its calibrated variant. In addition, we test our proposed Universal method with different numbers of images (K=3, 6, 15, 30, and 96). With as few as 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 6. On this dataset, we can see that our Universal method still achieves state-of-the-art results. It improves the state-of-the-art results by 13%, from 14.96° to 13.19°. On difficult geometries, like Turbine, the improvement is significant, and we also obtain good results on specular material like aluminium (AL), brass (CU) or steel. Our transformer calibrated method is however the best performer on this dataset (see Table 7). 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.

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

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 7: Mean angular error (in degrees, lower is better) 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. 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

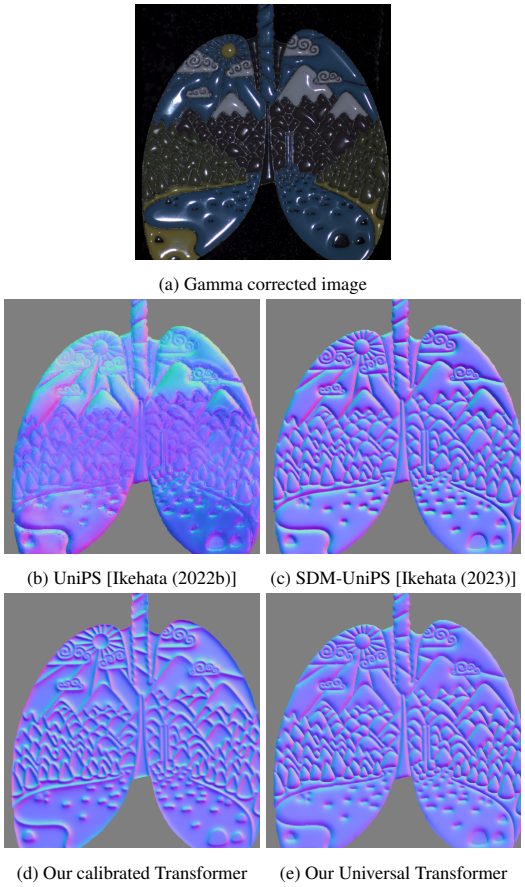
Table 8: Mean angular error (in degrees) on the DiLiGenT-Pi benchmark [Wang et al. (2023)]. Best results are in bold, and the second best ones are underlined. The type C means calibrated PS, UC is uncalibrated PS and Uni is Universal PS. The proposed method gives best state-of-the-art results.

		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. (2023)]	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
UniPS [Ikehata (2022b)]	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 9: Mean angular error (in degrees) on the Luces benchmark [Mecca et al. (2021)]. The proposed method provides the best results among uncalibrated 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 8. Compared to the calibrated methods, our Universal method tends to have slightly lower performance on near-flat objects, but it still 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 Fig. 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 Fig. 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 Fig. 10, our uncalibrated methods is able to predict correctly the normal map contrary to SDM-UniPS [Ikehata (2023)] and UniPS [Ikehata (2022b)]. 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 also able to manage non-parallel light beam as shown in Table 9. It reaches results close to the best calibrated results Lichy et al. (2022), yet our proposed approach relaxes the need for the tedious calibration of the various illumination parameters (location, intensities, anisotropy, etc.) involved in calibrated methods.



(a) Gamma corrected image

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

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

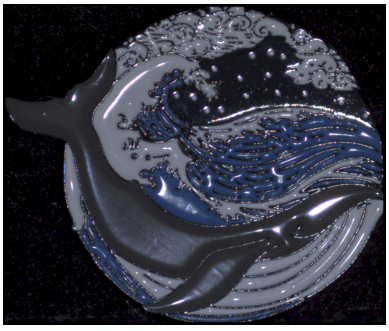
Fig. 9: Normal prediction of our methods, SDM-UniPS [Ikehata (2023)] and UniPS [Ikehata (2022b)] on the Lung object of Wang et al. (2023). We can see that this material is challenging for uncalibrated and Universal approaches because of the light reflection. Indeed, normal maps are inverted.

		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

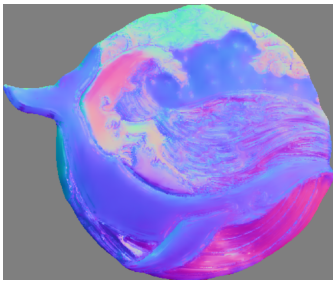
Table 10: Mean angular error (in degrees) on the DiLiGenT benchmark [Shi et al. (2016)] without masking the background before processing. For comparison the results with background is also shown.

		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

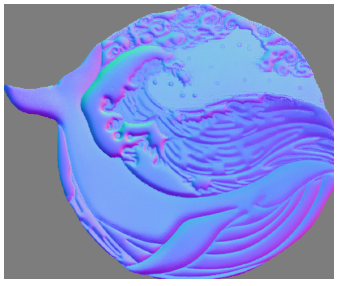
Table 11: Mean angular error (in degrees) on the Luces benchmark [Mecca et al. (2021)] with and without masking the background before processing.



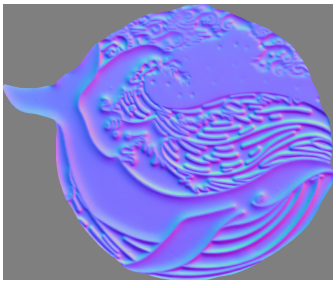
(a) Gamma corrected image



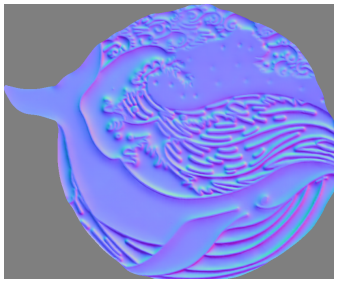
(b) UniPS [Ikehata (2022b)]



(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 UniPS [Ikehata (2022b)] 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.

4.3. 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 Tables 10 and 11), 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 Fig. 11, 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. 11: Comparison on the Alligator object of SDM-UniPS [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, and that we reconstruct more accurate details than SDM-UniPS [Ikehata (2023)].

4.4. 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, except UniPS [Ikehata (2022b)] which is known to be less accurate than SDM-UniPS. 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 12 and 13). In the Owl object in Fig. 12, 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 Fig. 14, the images resolution are really high and our Universal method obtains excellent results. Our method performs better than SDM-UniPS [Ikehata (2023)]. Although SDM-UniPS should in theory be used directly on the whole image, an alternative to manage really high resolution images could be to divide the whole images in patches (1024 by 1024 pixels by example) with overlapping pixels. As shown in Fig. 14, such an ad hoc procedure fails on objects with light interaction on the whole image, which further highlights the benefits of the multi-scale approach.

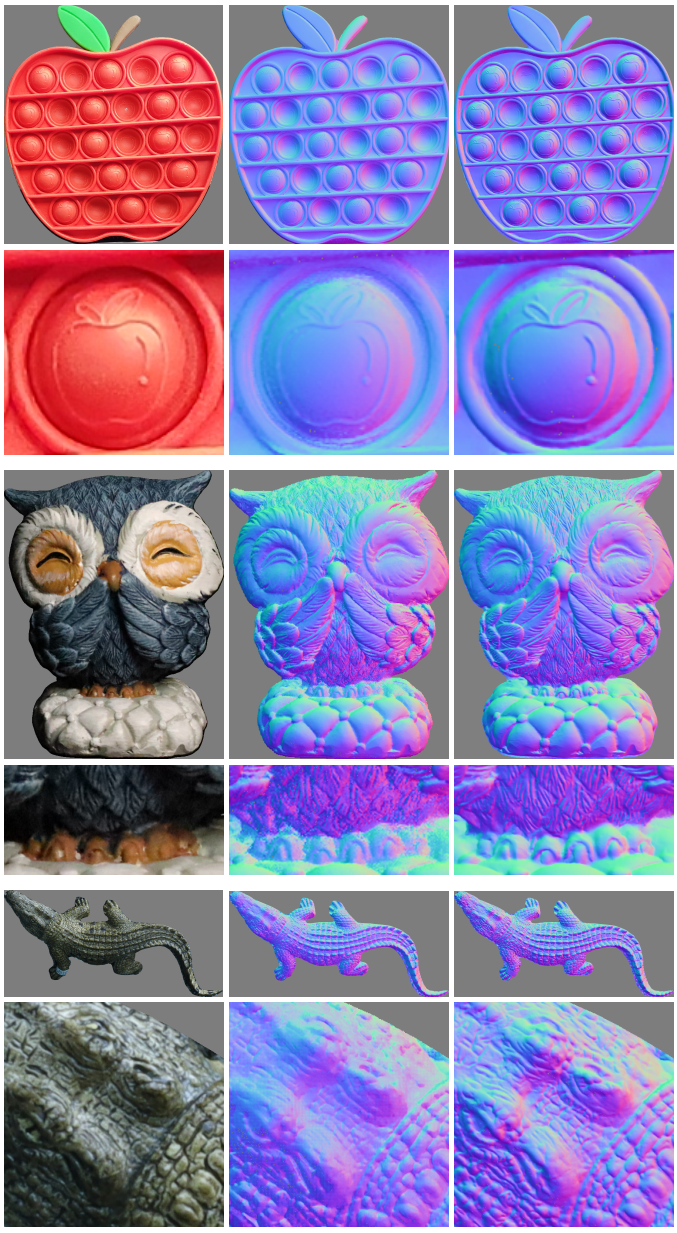
4.5. Limitations

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 Fig. 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, and further experiments on specifically designed datasets such as [Guo et al. (2024)] would be worthwile.

5. Conclusion

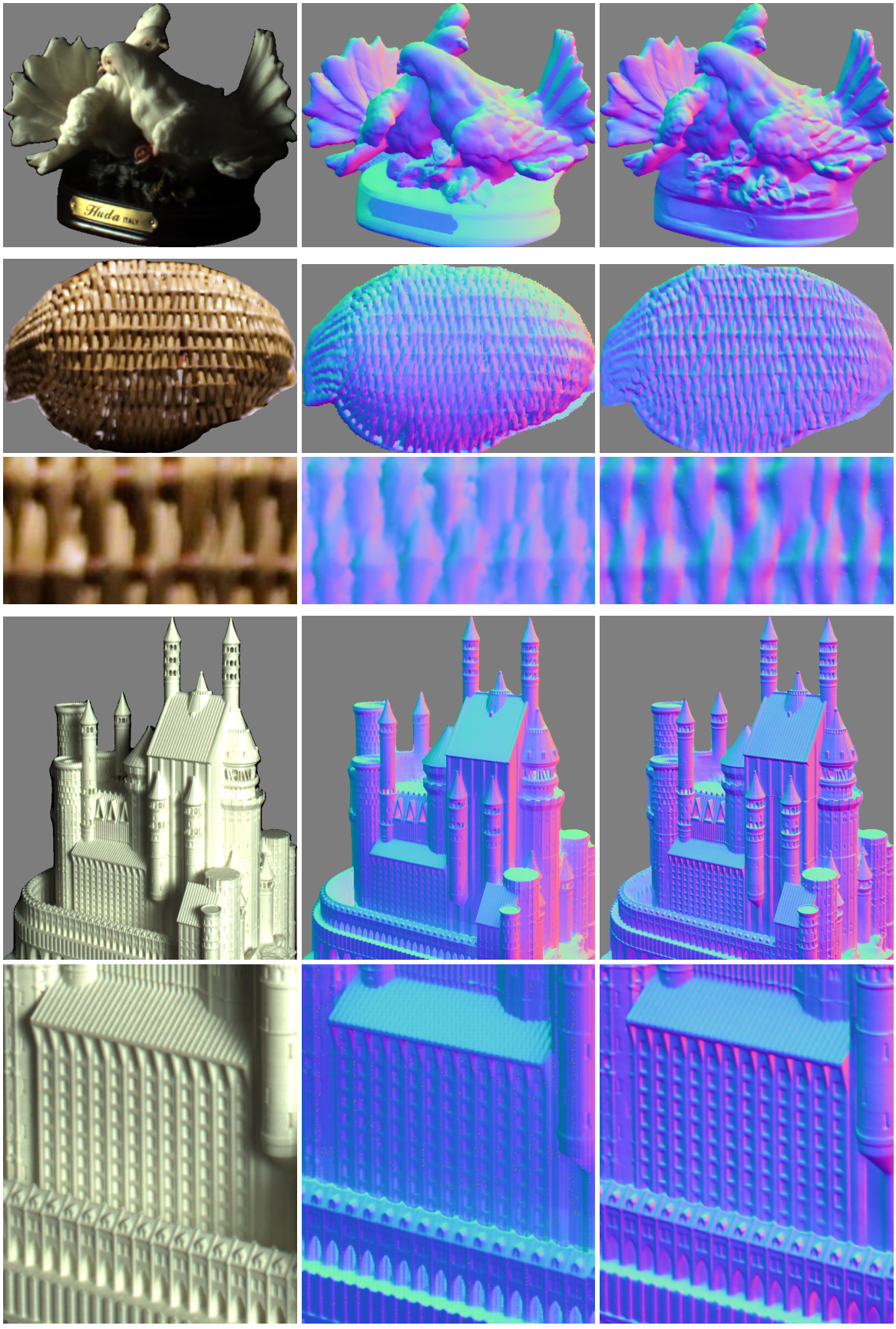
To conclude, we proposed a new multi-scale approach based on Transformers with encoder and decoder for each scale. Our method gives excellent results over a large panel of benchmark datasets with a large diversity of acquisition setups and environments, which show its robustness. Our method also shows 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.



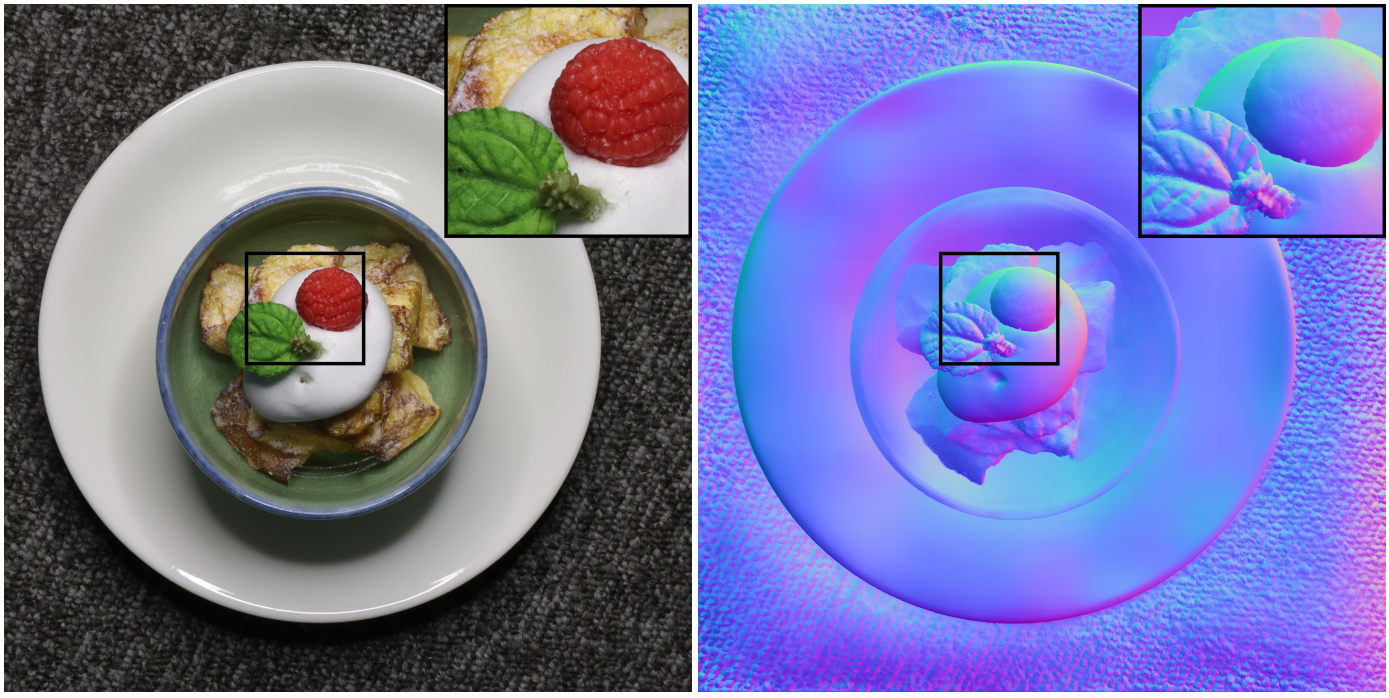
SDM-UniPS [Ikehata (2023)] SDM-UniPS [Ikehata (2023)] Our Universal Transformer

Fig. 12: Comparison on objects without ground truth from Ikehata (2022b, 2023). The first column is the RGB images, the second one is the SDM-UniPS method [Ikehata (2023)] 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-UniPS [Ikehata (2023)].

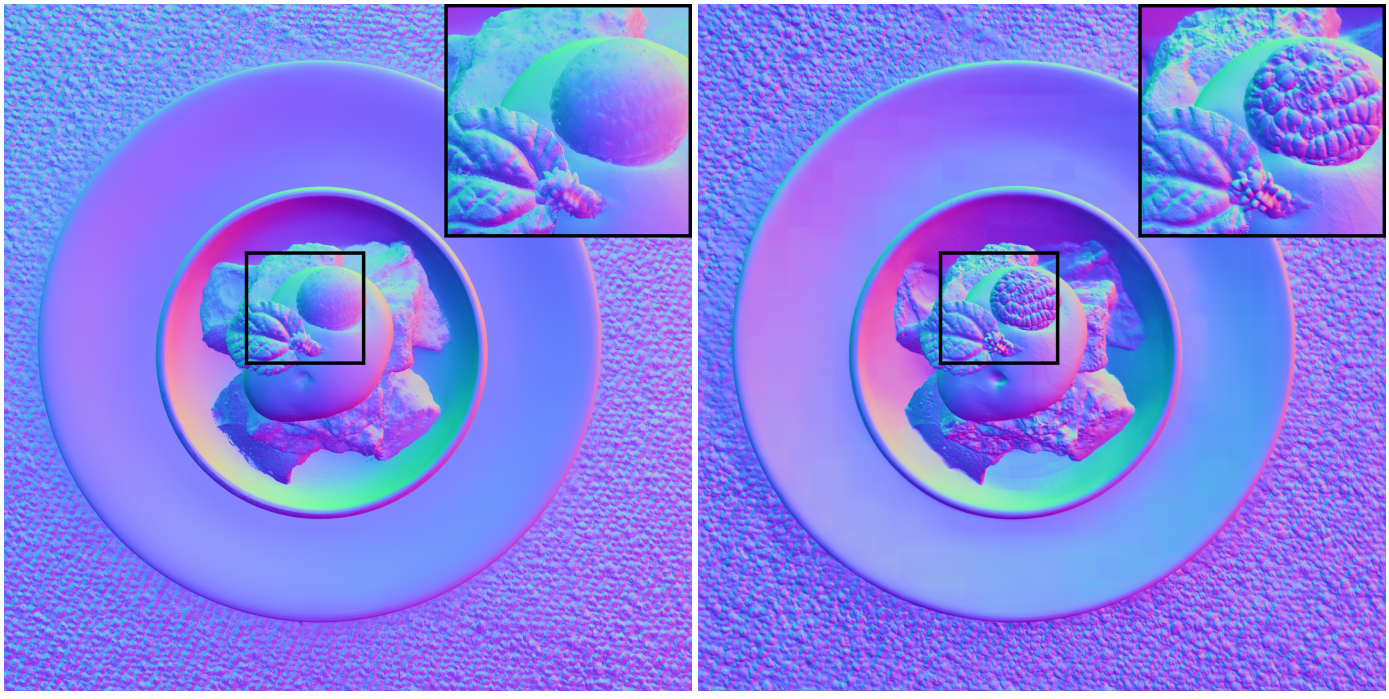


SDM-UniPS [Ikehata (2023)] SDM-UniPS [Ikehata (2023)] Our Universal Transformer

Fig. 13: 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 (2023)] 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 (2023)].



SDM-UniPS + patch-based method [Ikehata (2023)]



SDM-UniPS [Ikehata (2023)] Our

Fig. 14: Visual comparison beetween the 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

Chen, G., Han, K., S., B., M., Y., W., K., 2022. Deep Photometric Stereo for Non-Lambertian Surfaces. PAMI 44.

Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.Y.K., 2019. Self-Calibrating Deep Photometric Stereo Networks, in: CVPR.

Chen, G., Han, K., Wong, K., 2018. PS-FCN: A Flexible Learning Framework for Photometric Stereo, in: ECCV.

Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A., 2018. Single-image svbrdf capture with a rendering-aware deep network. SIGGRAPH 37, 15.

Guo, H., Ren, J., Wang, F., Shi, B., Ren, M., Matsushita, M., 2024. DiLiGenRT: A Photometric Stereo Dataset with Quantified Roughness and Translucency, in: CVPR, pp. 11810–11820.

Haefner, B., Ye, Z., Gao, M., Wu, T., Quéau, Y., Cremers, D., 2019. Variational uncalibrated photometric stereo under general lighting, in: ICCV, pp. 8539–8548.

Hardy, C., Queau, Y., Tschumperle, D., 2023. MS-PS: A Multi-Scale Network for Photometric Stereo With a New Comprehensive Training Dataset. Computer Science Research Notes 3301, 194–203.

Honzátko, D., Türetken, E., Fua, P., Dunbar, L., 2021. Leveraging Spatial and Photometric Context for Calibrated Non-Lambertian Photometric Stereo, in: 3DV.

Ikehata, S., 2018. CNN-PS: CNN-based Photometric Stereo for General Non-Convex Surfaces, in: ECCV.

Ikehata, S., 2021. PS-transformer: Learning sparse photometric stereo network using self-attention mechanism, in: BMVC.

Ikehata, S., 2022a. Does Physical Interpretability of Observation Map Improve Photometric Stereo Networks?, in: ICIP.

Ikehata, S., 2022b. Universal photometric stereo network using global lighting contexts. CVPR .

Ikehata, S., 2023. Scalable, detailed and mask-free universal photometric stereo, in: CVPR.

Ju, Y., Dong, J., Chen, S., 2021. Recovering Surface Normal and Arbitrary Images: A Dual Regression Network for Photometric Stereo. TIP 30, 3676–3690.

Ju, Y., Lam, K., Chen, Y., Qi, L., Dong, J., 2020. Pay Attention to Devils: A Photometric Stereo Network for Better Details, in: IJCAI.

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. 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: CVPR.

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: ICML, pp. 3744–3753.

Li, J., Li, H., 2022. Self-calibrating photometric stereo by neural inverse rendering, in: ECCV, pp. 166–183.

Li, J., Robles-Kelly, A., You, S., Matsushita, Y., 2019. Learning to Minify Photometric Stereo, in: CVPR.

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: CVPR.

Lichy, D., Sengupta, S., Jacobs, D., 2022. Fast light-weight near-field photometric stereo, in: CVPR.

Lichy, D., Wu, J., Sengupta, S., Jacobs, D., 2021. Shape and Material Capture at Home, in: CVPR.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. CVPR .

Logothetis, F., Budvytis, I., Mecca, R., Cipolla, R., 2021. PX-net: Simple and efficient pixel-wise training of photometric stereo networks, in: ICCV.

Logothetis, F., Mecca, R., Budvytis, I., Cipolla, R., 2023. A CNN based approach for the point-light photometric stereo problem. IJCV 131, 101–120.

Lorensen, W., Cline, H., 1987. Marching cubes: A high resolution 3D surface construction algorithm. SIGGRAPH .

Mecca, R., Logothetis, F., Budvytis, I., Cipolla, R., 2021. LUCES: A dataset for near-field point light source photometric stereo abs/2104.13135.

Mo, Z., Shi, B., Lu, F., Yeung, S.K., Matsushita, Y., 2018. Uncalibrated photometric stereo under natural illumination, in: CVPR, pp. 2936–2945.

Ren, J., Wang, F., Zhang, J., Zheng, Q., Ren, M., Shi, B., 2022. DiLiGenT10^2: A Photometric Stereo Benchmark Dataset with Controlled Shape and Material Variation, in: CVPR.

Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y., 2017. Deep Photometric Stereo Network, in: ICCV Workshops.

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: CVPR.

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: CVPR.

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: ICCV, pp. 9477–9487.

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: ICCV, pp. 568–578.

Wang, X., Jian, Z., Ren, M., 2020. Non-Lambertian Photometric Stereo Network Based on Inverse Reflectance Model With Collocated Light. TIP 29.

Woodham, R.J., 1980. Photometric Method For Determining Surface Orientation From Multiple Images. Opt. Eng. 19.

Yao, Z., Li, K., Fu, Y., Hu, H., Shi, B., 2020. GPS-Net: Graph-based Photometric Stereo Network, in: NIPS.

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: ICCV.