XII Jornada de Jóvenes Investigadores

In-the-wild Material Appearance Editing using Perceptual Attributes

FULL PAPER

J. Daniel Subias[1], Manuel Lagunas [2]

[1] Universidad de Zaragoza - I3A, Zaragoza, Spain

[2] Amazon

PROBLEM

Intuitively editing the appearance of materials, just from a single image, is a challenging task given the complexity and ambiguity of the interactions between light and matter.

This problem has been traditionally solved by estimating additional factors of the scene like geometry or illumination, thus solving an inverse rendering problem where the interaction of light and matter needs to be modelled.

We present a single-image appearance editing framework that allows to intuitively modify the material appearance of an object by increasing or decreasing high-level perceptual attributes describing appearance (e.g., glossy or metallic). Our framework uses just an in-the-wild image as input, where geometry or illumination are not controlled.

OUR FRAMEWORK

We introduce a novel framework that relies on an encoder-decoder architecture *G* that encodes the image **x**, and manipulates the latent space **z** together with the target attribute $\mathbf{att_t}$ to generate the edited image **y**. A high-level overview of our framework is shown in Figure 1. The STU architecture, illustrated in Figure 1, is a variant of the GRU [1,2] and allows encoder-decoder architectures to keep the relevant information of the input image in the edited output when manipulating the latent space **z**. Given the feature map of the l^{th} encoder layer denoted by $\mathbf{f}^l_{enc}$, the STU cell outputs an edited feature map $\mathbf{f}_t^l$, as is shown in Figure 1. Each STU cell receives information from the previous cell via a feature map $\hat{\mathbf{s}}^{l+1}$ (also called hidden state), which also contains information of the target attribute attt . The STU updates its internal hidden state denoted as sl and sends this to the next STU cell.

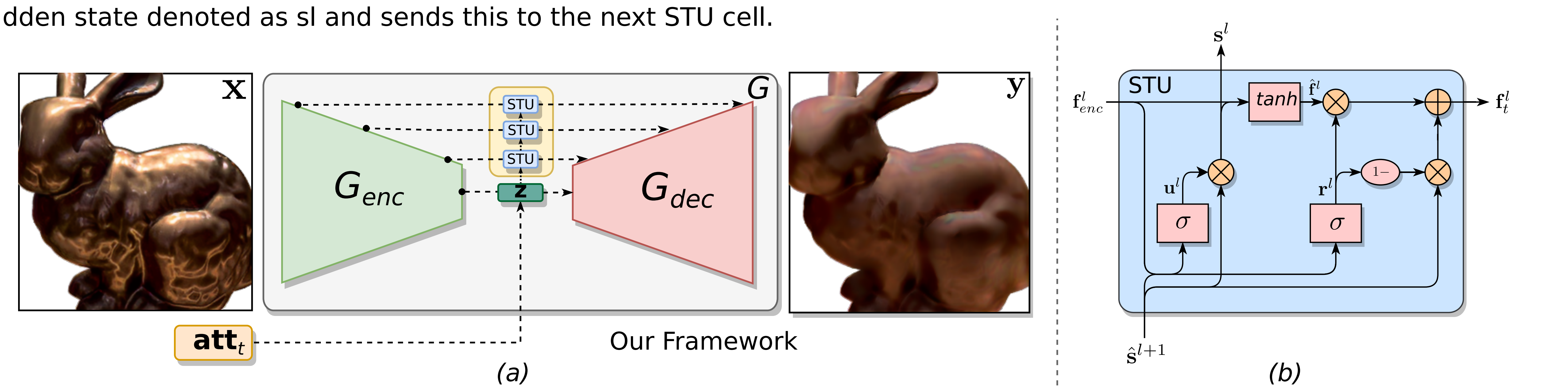

Figure 1: (a) High-level overview of our framework. Our generator *G* is composed of an encoder G_{enc} and a decoder G_{dec} . It is capable of editing the input image **x** according to the target attribute $\mathbf{att_t}$ to generate the edited image **y**. (b) The architecture of a single STU cell. As an input, it takes the feature map of the current layer $\mathbf{f}^l_{enc}$ and the hidden state of the previous cell $\hat{\mathbf{s}}^{l+1}$. It outputs the updated hidden state $\mathbf{s}^l$ and feature map $\mathbf{f}_t^l$.

TRAINING SCHEME

We adopt the adversarial training proposed by He et al. [3] and introduce a GAN model where the discriminator *D* has two branches D_{adv} and D_{att} . D_{adv} consists of five convolution layers to predict whether an image is fake (edited) or real. D_{att} shares the convolution layers with D_{adv} and, instead, predicts the high-level attribute value $\boldsymbol{att_t}$. Figure 2 shows a high-level scheme of our framework during training.

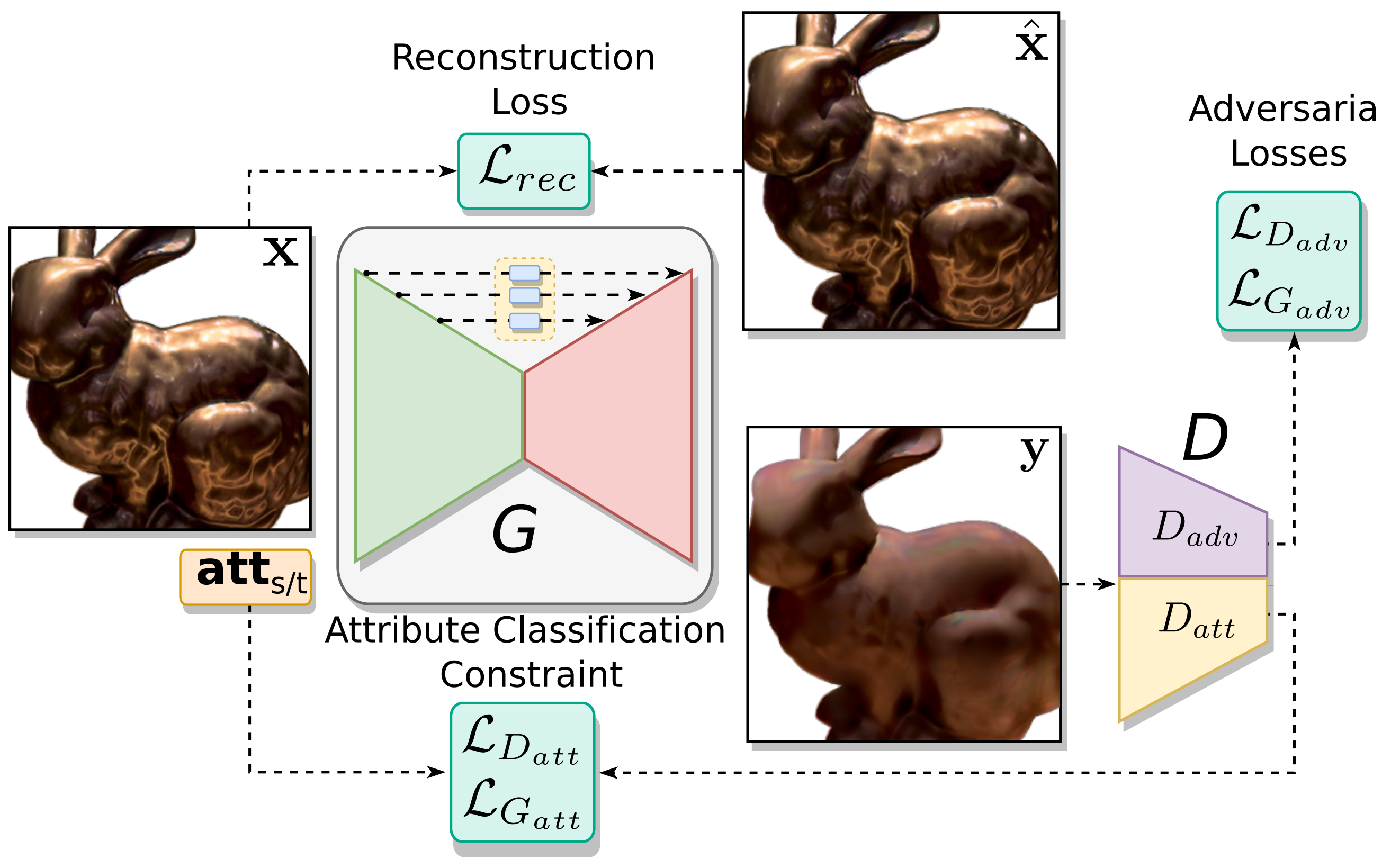

Figure 2: Training scheme of our framework. The gray block represents our generator G, and the discriminator branches D_{adv} and D_{att} are illustrated by the purple and yellow blocks, respectively. Pointed arrows denote the parameters used as input for the training

We leverage the dataset of Delanoy et. al [4], designed for material appearance perception tasks. The dataset has 45,500 images of single-object scenes. (13 geometries × 100 materials × 7 illuminations × 5 views). Figure 3 shows few examples of images present in the dataset.

Figure 3: Five scenes of the geometries present in the training dataset with different materials under different illuminations.

RESULTS

Our approach learns to edit perceptual cues properly while objects' shape remains unchanged (Figure 4). We compare our results against the method of Delanoy et al. [4]. We show results editing the input using a different target perceptual attribute.

Since the method of Delanoy et al. also needs the normal map as the input, we evaluate in-the-wild photographs with their estimated normal map (using Delanoy et al. estimator), and synthetic images with a perfectly normal map. Our method does not need the normal map as the input.

In Figure 4 we can see a comparison between the edited images by our method and the one by Delanoy et al. [4]. Our approach learns to edit perceptual cues properly while objects' shape remains unchanged. The material appearance edits from Delanoy et al. [4] strongly depend on the shape of their estimated normal map [4]. This causes geometry details that are not present in the normal map not to be present in the edited image.

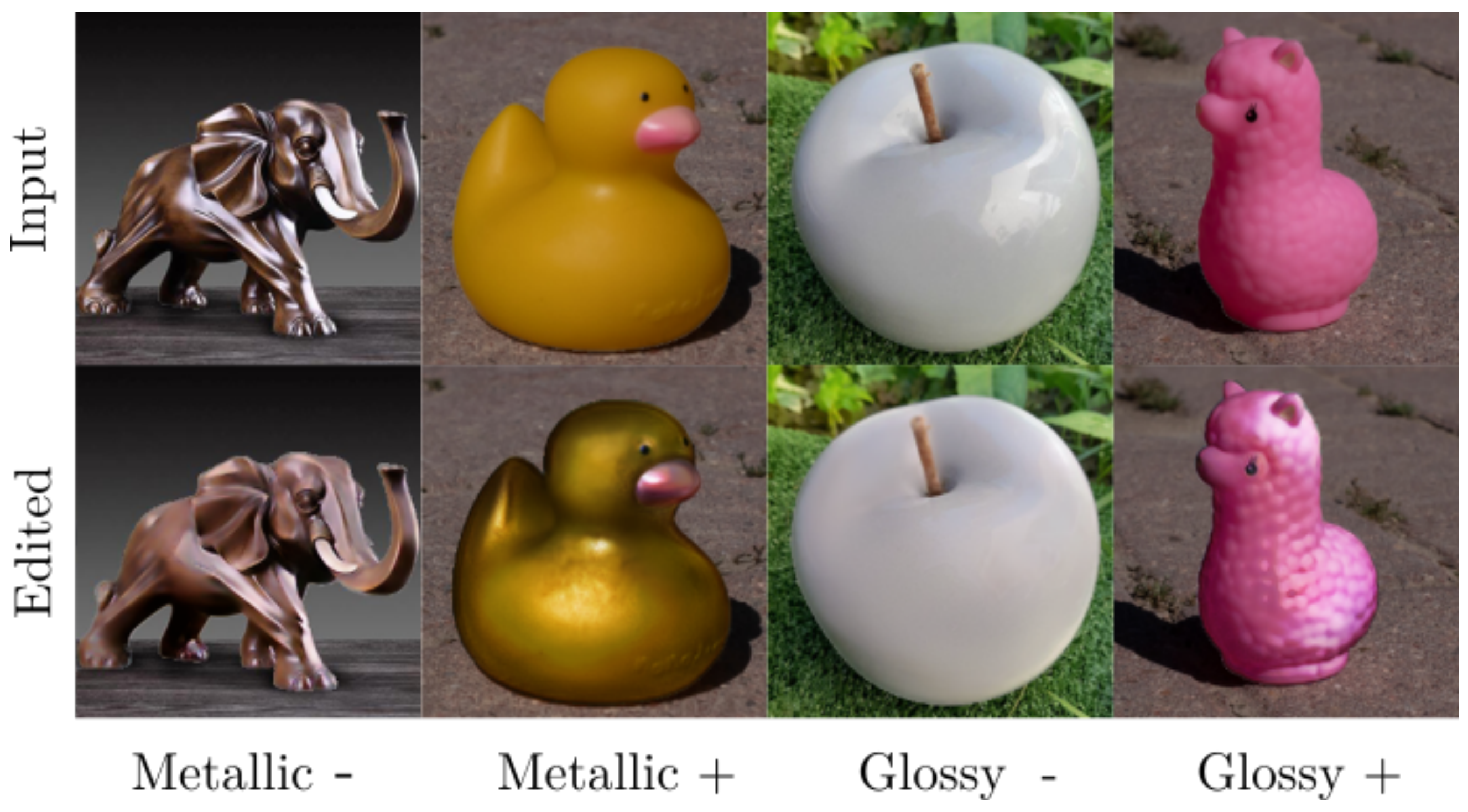

Figure 4: A sample of the real photographs edited by our framework without suplementary information of the scene. The "+" and "-" indicate whether the target high-level perceptual attribute increased or decreased.

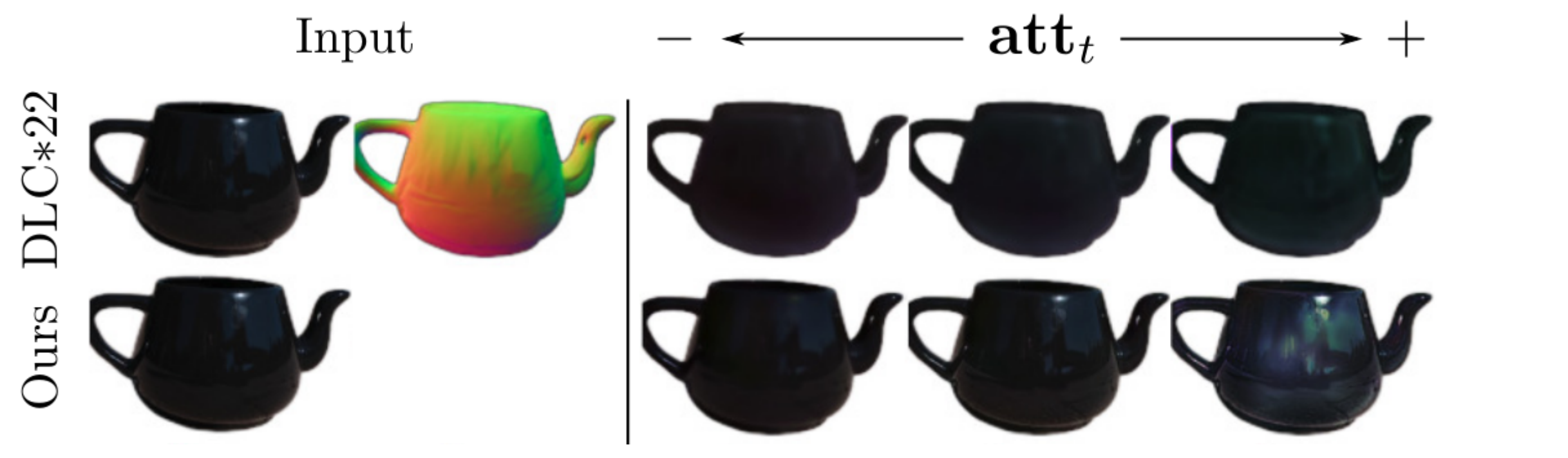

Figure 5: Comparison editing the glossy attribute using the method of Delanoy et al. [4] and our framework for two in-the-wild photographs. Our framework only requires the photograph as the input while the work of Delanoy et al. needs to estimate the normal map. We can see how our method better recovers the glossy appearance of the object when edited. Besides, it is able to recover better high-frequency details. The "+" and "-" indicate whether the target high-level perceptual attribute increased or decreased.

REFERENCES

[1]- CHO K., VAN MERRIËNBOER B., GULCEHRE C., BAH-DANAU D., BOUGARES F., SCHWENK H., B ENGIO Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proc. Empirical Methods in Natural Language Processing (EMNLP) (Doha, Qatar, Oct. 2014), Association for Computational Linguistics, pp. 1724–1734.

[2]- CHUNG J., GÜLÇEHRE Ç., CHO K., BENGIO Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling.CoRR abs/1412.3555 (2014).

[3]- HE Z., ZUO W., KAN M., SHAN S., CHEN X.: Attgan: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing 28, 11 (Nov 2019), 5464–5478.

[4]- DELANOY J., LAGUNAS M., CONDOR J., GUTIERREZ D., MASIA B.: A generative framework for image-based editing of material appearance using perceptual attributes. Computer Graphics Forum 41, 1(2022), 453–464.

ACKNOWLEDGEMENTS

This project has received funding from the Government of Aragon's Departamento de Ciencia, Universidad y Sociedad del Conocimiento through the Reference Research Group "Graphics and Imaging Lab", the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956585 (PRIME), the CHAMELEON project (European Union's Horizon 2020, European Research Council, grant agreement No. 682080), and MCIN/AEI 10.13039/501100011033 through Project PID2019-105004GB-I00.

PRIME
GOBIERNO DE ARAGON
erc European Research Council
Established by the European Commission
AGENCIA ESTATAL DE INVESTIGACIÓN

Graphics and Imaging Lab
amazon
Instituto Universitario de Investigación
de Ingeniería de Aragón
Universidad Zaragoza

Contact: dsubias@unizar.es