Intrinsic Autoencoders for Joint Neural Rendering and Intrinsic Image Decomposition - Supplementary materials.

Hassan Abu Alhaija[*,1] Siva Karthik Mustikovela[*,1] Justus Thies[2]
Varun Jampani[3] Matthias Nießner[2] Andreas Geiger[4,5] Carsten Rother[1]

[1]Heidelberg University [2]Technical University Munich [3]Google Research
[4]Max Planck Institute for Intelligent Systems, Tübingen [5]University of Tübingen

1. Overview

In this supplement, we provide the architectural and training details of our framework in sections. We also provide several qualitative results for neural deferred rendering and intrinsic image decomposition.
In section 2, we describe the architecture of our neural deferred renderer ($\mathcal{R}$), intrinsic image decomposition network ($\mathcal{H}$), image discriminator $\mathcal{D}_I$ and intrinsics discriminator $\mathcal{D}_M$. In section 3, we present various training hyperparameters and training schedule. Section 4.1 contains qualitative comparisons between the rendering produced by our method and those produced by baselines. Section 4.2 contains qualitative comparisons between the intrinsic decompositions produced by our method and those produced by baselines. We will release our code upon publication

2. Network Architectures

This section provides the architectures for each of the networks used in our work.

2.1. Neural Deferred Renderer

The input to this network consists of albedo ($H \times W \times 3$), normals ($H \times W \times 3$) and reflections ($H \times W \times 3$) concatenated into a single data block of size $H \times W \times 9$, where H is the image height of 256 and W is image width of 512 To achieve high resolution image synthesis, we use the coarse-to-fine generator introduced in [4]. The input first goes through a set of convolutions with a kernel size 7×7 and 512 channels with stride 1 followed by two convolutional layers with a kernel size of 3×3 each with 512 channels and stride of 2. Next, we use a series of 9 ResNet blocks each with a kernel size 3×3 and 256 channels. Finally, we use two transpose convolutional layers to upsample the features. The first has a kernel size 3×3 and stride 2. This is followed by a transpose convolutional layer with kernel size 7×7 and stride 1 which produces 3 output channels. The final output of the generator is a 3 channel RGB image $\hat{I}$ of size $256 \times 512 \times 3$.

2.2. Intrinsic Image Decomposition Networks

We use three networks $\mathcal{H} = \{\mathcal{H}_N, \mathcal{H}_A, \mathcal{H}_F\}$ for estimating the surface normals N, Albedo A and environment reflections F, respectively, from an image I. The input image first goes through a set of convolutions with a kernel size 5×5 with a stride 1 and 512 channels followed by two convolution layers of size 3×3 with stride 2 and 256 channels. This is followed by 5 Resnet blocks each with a kernel size 3×3. Finally, we use transpose convolutional layers with a kernel size 3×3 with stride 1 to upsample these features. The last transpose convolution layer has a kernel size of 7×7 and stride 1 which produces output of size $256 \times 512 \times 3$. All networks in $\mathcal{H}$ have the same architecture described above.

2.3. Discriminator networks

The image discriminator network takes an RGB image as input. We use a multi-scale PatchGAN discriminator [4] which comprises two fully-convolutional networks that classify the local image patches. The first operates on the full resolution of the image and the second operates on the image downscaled by a factor of 2. Each discriminator network consists of 4 convolutional layers each with kernel size 4×4, stride 2 and filter numbers of 64, 128, 256 and 512. At the end, a convolutional layer with kernel size 1×1 and stride 1 combines the features into a 1 channel output. The discriminators output a realism score for each patch instead of a single prediction per image. The intrinsics discriminator $\mathcal{D}_M$ has the same architecture except that the input is a 9-channel stack combining all three intrinsic maps.

3. Training Details

Our code is implemented in Pytorch [3]. We will release all code and training models required to reproduce the results. We train all our networks from scratch by initializing all weights with a normal distributionN(0,0.2)and zero bias. The learning rate for all networks is 0.0001. We use Adam[2] optimizer with betas (0.9, 0.99) and no weight decay. We train the networks for 30 epochs.

4. Visual Results

In this section we qualitatively compare our results to other baselines for the task of Neural Deferred Rendering and Intrinsic Image Decomposition. We compare our results to CycleGAN[5], MUNIT[1] and our own baselines.

4.1. Neural Deferred Rendering Results

Figures 1, 2, 3, 4, 5 illustrate renderings produced by our model compared to those produced by other baseline methods.

4.2. Intrinsic Image Decomposition Results

Figures 6, 7, 8, 9 illustrate intrinsic image decompositions produced by our model compared to those produced by other baseline methods.

References

[1] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 2

[2] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 1

[3] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 1

[4] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 1

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 2

Figure 1: **Qualitative Comparison with baselines on Neural Rendering.** Inputs to the network illustrated in top left are intrinsic maps consisting of normals, albedo and reflections. *Our full model* produces highly photorealistic images which preserve input geometry, albedo and reflections. The model trained without shared discriminator (*w/o shared discr.*), produces images with significant artefacts especially near the wheels. The model trained without decomposition cycle (*w/o* Decomp cycle), can produce inconsistent images that don't match the input normals, albedo. Both *CycleGAN* and *MUNIT* produce significantly worse images with strong artifacts compared to our model.

Figure 2: **Qualitative Comparison with baselines on Neural Rendering.** Inputs to the network illustrated in top left are intrinsic maps consisting of normals, albedo and reflections. *Our full model* produces highly photorealistic images which preserve input geometry, albedo and reflections. The model trained without shared discriminator (*w/o shared discr.*), produces images with significant artefacts especially near the wheels. The model trained without decomposition cycle (*w/o* Decomp cycle), can produce inconsistent images that don't match the input normals, albedo. Both *CycleGAN* and *MUNIT* produce significantly worse images with strong artifacts compared to our model.

Figure 3: **Qualitative Comparison with baselines on Neural Rendering.** Inputs to the network illustrated in top left are intrinsic maps consisting of normals, albedo and reflections. *Our full model* produces highly photorealistic images which preserve input geometry, albedo and reflections. The model trained without shared discriminator (*w/o shared discr.*), produces images with significant artefacts especially near the wheels. The model trained without decomposition cycle (*w/o* Decomp cycle), can produce inconsistent images that don't match the input normals, albedo. Both *CycleGAN* and *MUNIT* produce significantly worse images with strong artifacts compared to our model.

Figure 4: **Qualitative Comparison with baselines on Neural Rendering.** Inputs to the network illustrated in top left are intrinsic maps consisting of normals, albedo and reflections. *Our full model* produces highly photorealistic images which preserve input geometry, albedo and reflections. The model trained without shared discriminator (*w/o shared discr.*), produces images with significant artefacts especially near the wheels. The model trained without decomposition cycle (*w/o* Decomp cycle), can produce inconsistent images that don't match the input normals, albedo. Both *CycleGAN* and *MUNIT* produce significantly worse images with strong artifacts compared to our model.

Figure 5: **Qualitative Comparison with baselines on Neural Rendering.** Inputs to the network illustrated in top left are intrinsic maps consisting of normals, albedo and reflections. *Our full model* produces highly photorealistic images which preserve input geometry, albedo and reflections. The model trained without shared discriminator (*w/o shared discr.*), produces images with significant artefacts especially near the wheels. The model trained without decomposition cycle (*w/o* Decomp cycle), can produce inconsistent images that don't match the input normals, albedo. Both *CycleGAN* and *MUNIT* produce significantly worse images with strong artifacts compared to our model.

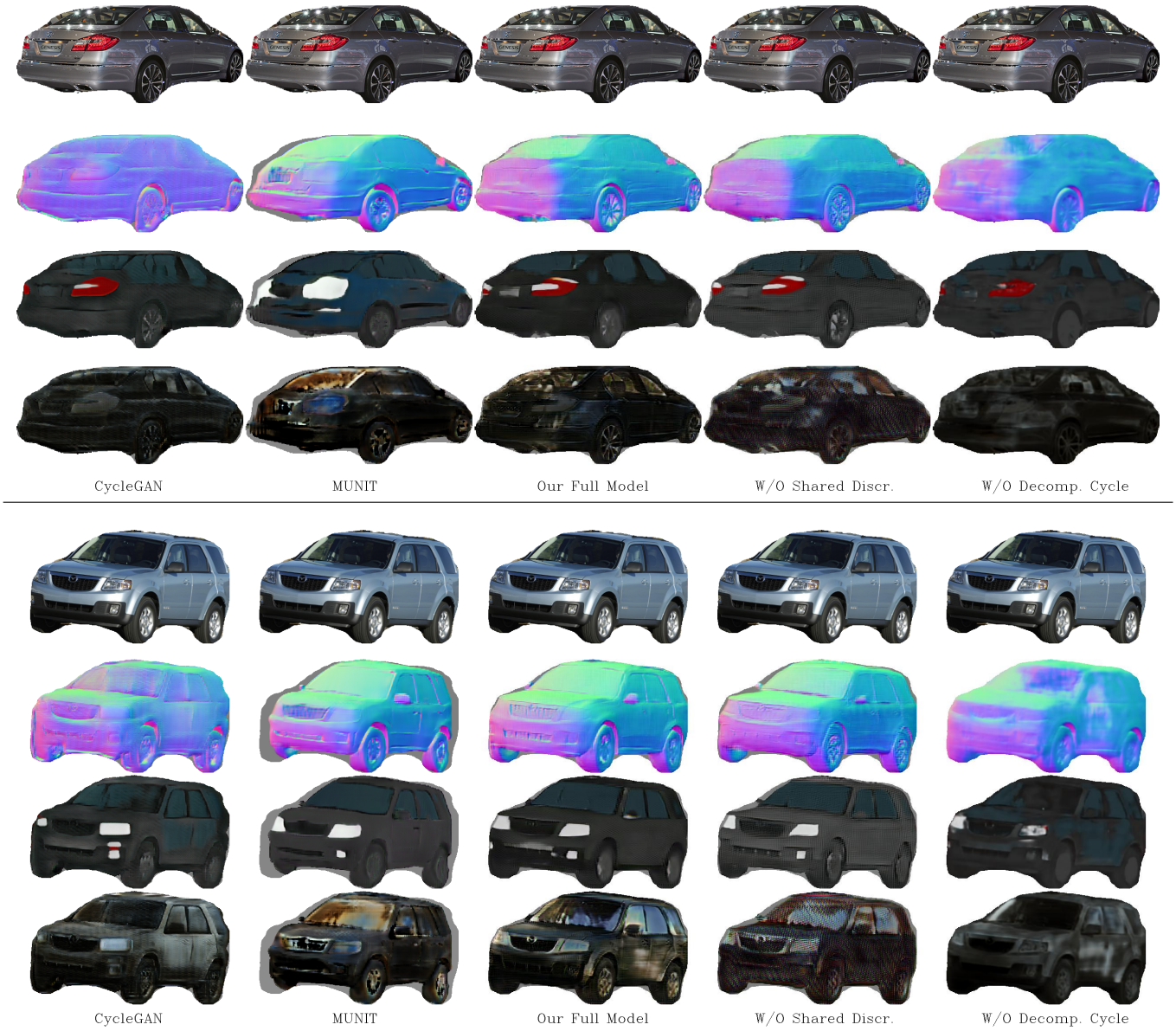

Figure 6: **Qualitative Comparison with baselines on Intrinsic Image Decomposition.** The top row indicates the real image input. The second row contains predicted normal maps. Third row contains predicted albedo. Last row contains predicted reflections. *CycleGAN* produces extremely noisy decompositions with artifacts. *MUNIT* predictions do not correspond well to the input image. Alongside, the reflections and albedo are noisy. *Our full model* is able to generalize to real world data and hence estimates accurate intrinsic maps. The model *w/o shared discr.* leads to high frequency artefacts in reflections and normals. Our model without decomposition cycle (*w/o Decomp. cycle*) also recovers noisy albedo and normals since the networks to overfit only to synthetic data leading to poor generalization to real images.

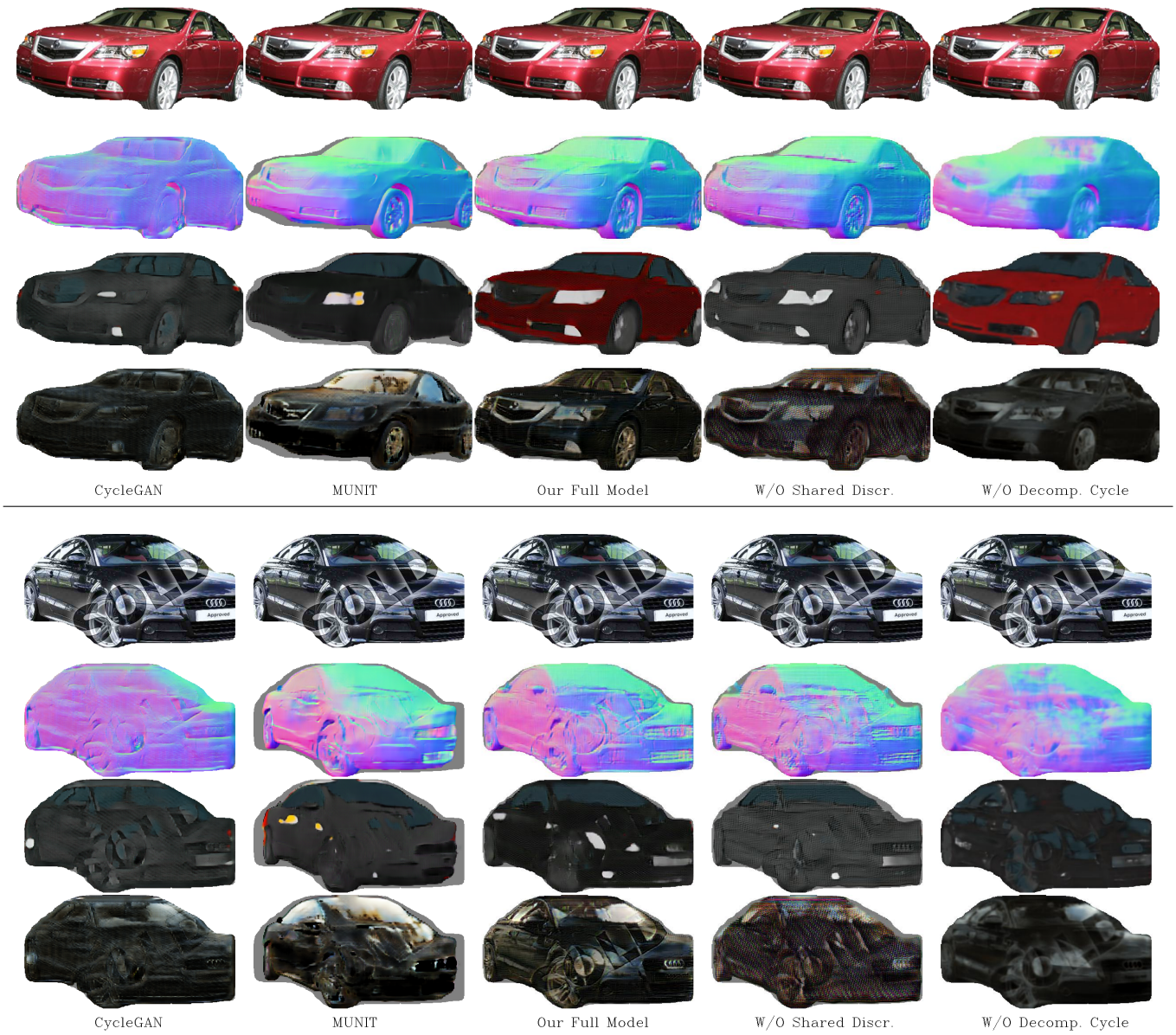

Figure 7: **Qualitative Comparison with baselines on Intrinsic Image Decomposition.** The top row indicates the real image input. The second row contains predicted normal maps. Third row contains predicted albedo. Last row contains predicted reflections. *CycleGAN* produces extremely noisy decompositions with artifacts. *MUNIT* predictions do not correspond well to the input image. Alongside, the reflections and albedo are noisy. *Our full model* is able to generalize to real world data and hence estimates accurate intrinsic maps. The model *w/o shared discr.* leads to high frequency artefacts in reflections and normals. Our model without decomposition cycle (*w/o Decomp. cycle*) also recovers noisy albedo and normals since the networks to overfit only to synthetic data leading to poor generalization to real images.

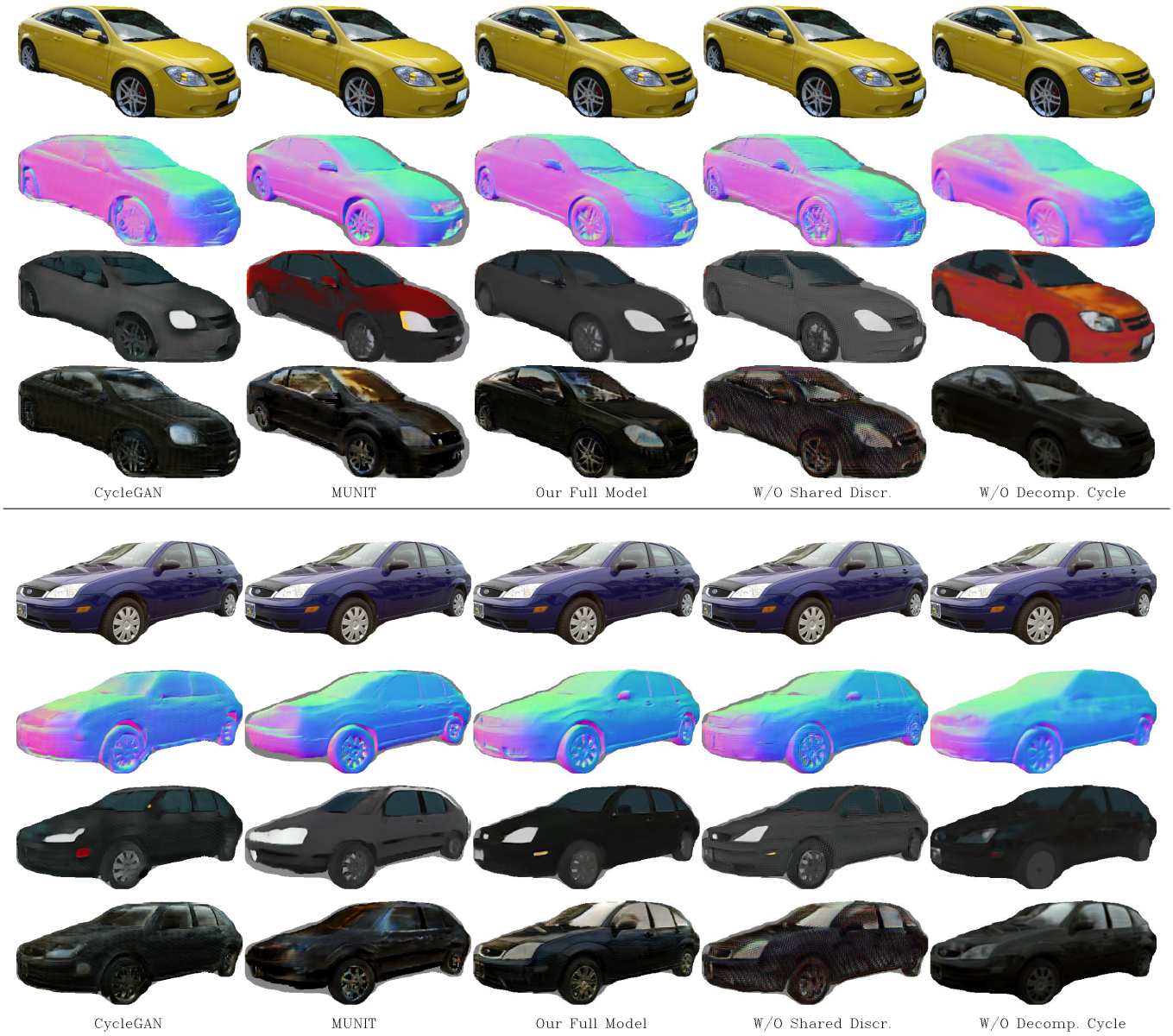

Figure 8: **Qualitative Comparison with baselines on Intrinsic Image Decomposition.** The top row indicates the real image input. The second row contains predicted normal maps. Third row contains predicted albedo. Last row contains predicted reflections. *CycleGAN* produces extremely noisy decompositions with artifacts. *MUNIT* predictions do not correspond well to the input image. Alongside, the reflections and albedo are noisy. *Our full model* is able to generalize to real world data and hence estimates accurate intrinsic maps. The model *w/o shared discr.* leads to high frequency artefacts in reflections and normals. Our model without decomposition cycle (*w/o Decomp. cycle*) also recovers noisy albedo and normals since the networks to overfit only to synthetic data leading to poor generalization to real images.

Figure 9: **Qualitative Comparison with baselines on Intrinsic Image Decomposition.** The top row indicates the real image input. The second row contains predicted normal maps. Third row contains predicted albedo. Last row contains predicted reflections. *CycleGAN* produces extremely noisy decompositions with artifacts. *MUNIT* predictions do not correspond well to the input image. Alongside, the reflections and albedo are noisy. *Our full model* is able to generalize to real world data and hence estimates accurate intrinsic maps. The model *w/o shared discr.* leads to high frequency artefacts in reflections and normals. Our model without decomposition cycle (*w/o Decomp. cycle*) also recovers noisy albedo and normals since the networks to overfit only to synthetic data leading to poor generalization to real images.

Figure 10: **Qualitative Comparison with baselines on Intrinsic Image Decomposition.** The top row indicates the real image input. The second row contains predicted normal maps. Third row contains predicted albedo. Last row contains predicted reflections. *CycleGAN* produces extremely noisy decompositions with artifacts. *MUNIT* predictions do not correspond well to the input image. Alongside, the reflections and albedo are noisy. *Our full model* is able to generalize to real world data and hence estimates accurate intrinsic maps. The model *w/o shared discr.* leads to high frequency artefacts in reflections and normals. Our model without decomposition cycle (*w/o Decomp. cycle*) also recovers noisy albedo and normals since the networks to overfit only to synthetic data leading to poor generalization to real images.