# Supplementary Material for Deep Image Harmonization by Bridging the Reality Gap

Junyan Cao#
Joy_C1@sjtu.edu.cn
Wenyan Cong#
wycong@utexas.edu
Li Niu*
ustcnewly@sjtu.edu.cn
Jianfu Zhang
c.sis@sjtu.edu.cn
Liqing Zhang
zhang-lq@cs.sjtu.edu.cn

MoE Key Lab of Artificial Intelligence
Shanghai Jiao Tong University
Shanghai, China

In the Supplementary file, we will first introduce more details about our contributed dataset RdHarmony in Section 1, and introduce the implementation details in Section 2. Then, we will ablate each component of the loss function, investigate the network design, and analyze the impact of different hyper-parameters in Section 3. We will explore the performance variance using different settings of training data in Section 4. Besides, we will exhibit the harmonization results of different methods on real test set $\mathcal{D}^{rl}_{te,b}$ from base categories in Section 5 and compare our method with the upper bound (row 13 in Table 2 in the main paper) in Section 6. We will introduce more details of user study conducted on 46 real-world composite images with human foregrounds and show example harmonization results of different methods in Section 7. Then, we will demonstrate the generalization of our CharmNet by conducting experiments on 5 more novel categories in Section 8. Finally, we discuss the limitations in Section 9.

## 1 Dataset Construction

We construct rendered image harmonization dataset RdHarmony that contains rendered images with foregrounds of "human" and 5 object categories ("bottle", "cake", "motorcycle", "cow", and "couch") from different super-categories. Taking category "human" as an example, we first introduce the data generation process in Section 1.1 and Section 1.2. Then, we discuss the category extension in Section 1.3 and illustrate the diversity in Section 1.4.

### 1.1 Ground-truth Rendered Image Generation

Generating a large number of rendered images with human foregrounds requires various 3D human characters placed in different 3D scenes. Considering the intra-category variance of

#Equal contribution. *Corresponding author.

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

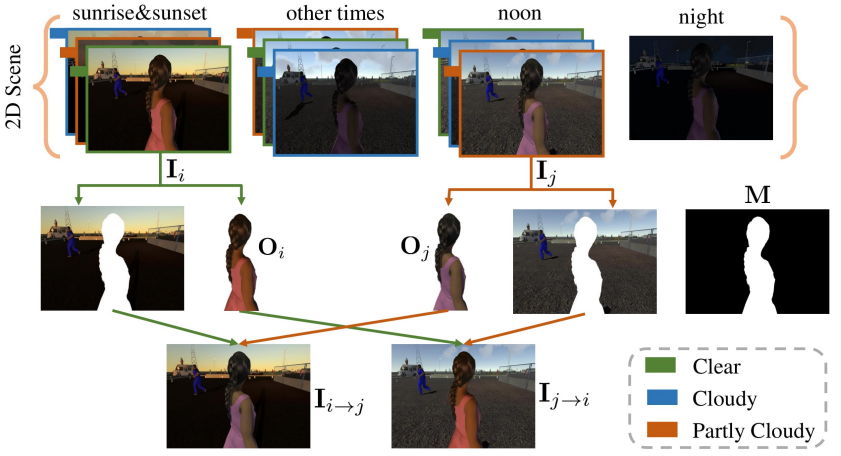



Figure 1: The generation process of composite rendered images. Top row: one 2D scene with 10 different styles. Middle row: exchange the foregrounds between two rendered images with different styles. Bottom row: obtained composite rendered images.

"human" category, we leverage the open-source software *MakeHuman* to create diverse 3D human characters with distinct attributes including skeleton, body features (*e.g.*, height, facial components), pose (*e.g.*, walking, running), and clothes. Actually, due to the prevalence of 3D modelling, there are abundant available 3D models of "human" category and other categories that could be used (see Section 1.3). Besides, we leverage the 3D game engine Unity3D and collect indoor and outdoor 3D scenes from Unity Asset Store and CG websites. Then, we import 3D human characters into the 3D scenes and vary the camera viewpoints to shoot various 2D scenes. In detail, we create 1,500 3D human characters, collect 30 3D scenes, and set 50 camera viewpoints for each 3D scene, leading to 1,500 2D scenes with unique foregrounds.

As claimed in Section 3 in the main paper, we employ UniStorm to control the weather and time while producing 2D scenes. Based on the different capture conditions, we define 10 representative styles, including the night style as well as styles of Clear/Partly Cloudy/Cloudy weather at sunrise&sunset/noon/other-times. For each 2D scene, we randomly sample one rendered image from each style, resulting in a group of 10 images. Given 1,500 2D scenes, we generate 15,000 rendered images with human foregrounds and different styles.

## 1.2 Composite Rendered Image Generation

As shown in Figure 1, for each 2D scene, we treat one person as foreground and obtain the foreground mask $\mathbf{M}$ effortlessly using Unity3D. We have 10 rendered images $\{\mathbf{I}_i|_{i=1}^{10}\}$ with different styles $\{\mathbf{y}_i|_{i=1}^{10}\}$ from the same 2D scene, where $\mathbf{y}_i$ is a 10-dim one-hot style label vector. We denote the foreground person in rendered image $\mathbf{I}_i$ as $\mathbf{O}_i$. After randomly selecting $\mathbf{I}_i$ and $\mathbf{I}_j$ from $\{\mathbf{I}_i|_{i=1}^{10}\}$, we could generate two pairs of composite rendered images and ground-truth rendered images $\{\mathbf{I}_{i\to j},\mathbf{I}_i\}$and $\{\mathbf{I}_{j\to i},\mathbf{I}_j\}$ by exchanging $\mathbf{O}_i$ and $\mathbf{O}_j$, where $\mathbf{I}_{j\to i}$ denotes the composite rendered image with foreground $\mathbf{O}_i$ and background of $\mathbf{I}_j$.

For rendered image pairs $\{\mathbf{I}_{j\to i},\mathbf{I}_j\}$, the style label of ground-truth rendered image $\mathbf{I}_j$ is $\mathbf{y}_j$ as mentioned above. For composite rendered image $\mathbf{I}_{j\to i}$, we simply assume that its style

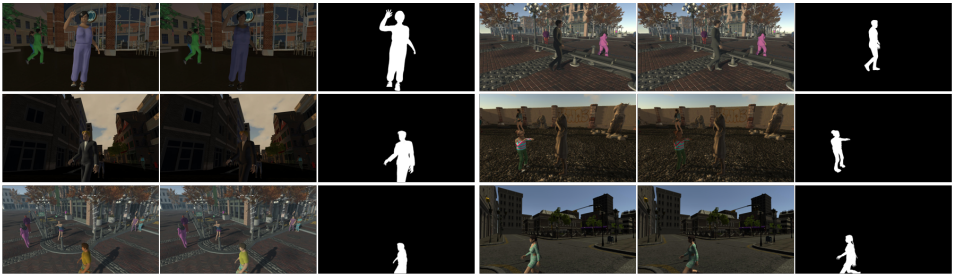



Figure 2: Examples of rendered training pairs from our RdHarmony with "human" category.

label vector is weighted average of foreground style $\mathbf{y}_i$ and background style $\mathbf{y}_j$ similar to [9]: $\mathbf{y}_{j\to i} = r \cdot \mathbf{y}_i + (1-r) \cdot \mathbf{y}_j$, in which foreground ratio $r$ denotes the area of foreground over the area of whole image.

We have generated 135,000 pairs of composite rendered images and ground-truth rendered images for "human" category in RdHarmony. Examples of rendered training pairs with "human" category are shown in Figure 2. It is worth mentioning that each above-mentioned step (*i.e.*, creating 3D characters, capturing 2D scenes, varying capture conditions, and generating composite rendered images) could be done automatically using scripts, making the dataset efficiently constructed and easily extendable. Besides, with controllable capture conditions (*i.e.*, styles) in UniStorm, it is feasible to acquire ground-truth style labels, which are useful for cross-domain knowledge transfer (Section 4.2 in the main text).

## 1.3 Considerable Extendability

We emphasize that we focus on human harmonization in this work, but the dataset construction could be easily generalized to other categories. *With the prosperity of Computer Graphics and photogrammetry, 3D model markets are common, making 3D models and 3D scenes more accessible than ever.* To be more specific, 3D models of certain categories could be automatically generated by specific software (*e.g.*, MakeHuman, SpeedTree). 3D models of other categories can be freely downloaded or purchased from Unity Asset Store and many public websites [1].

With 3D scenes and 3D models available, the rest of steps (*i.e.*, capturing 2D scenes, varying capture conditions, and generating composite rendered images) could be done automatically using scripts. To facilitate further study on image harmonization, all scrips used in our dataset construction will be made publicly available.

Furthermore, for better demonstration, we extend RdHarmony to 5 more categories from 5 super-categories, including "bottle", "cake", "motorcycle", "cow", and "couch". For each novel category, we place 3D models in each 3D scene, select one 3D model as the foreground and shoot a group of 2D scenes. In detail, for each category, we select 2 3D scenes, and set 50 camera viewpoints for each 3D scene, leading to 100 2D scenes. Then for each 2D scene, we obtain a group of 10 images with 10 different styles. For each category, we can generate 1,000 rendered images with different styles, and produce 9,000 pairs of composite rendered

[1] www.cgtrader.com; www.aigei.com; www.cgmodel.com; www.turbosquid.com; www.sketchfab.com

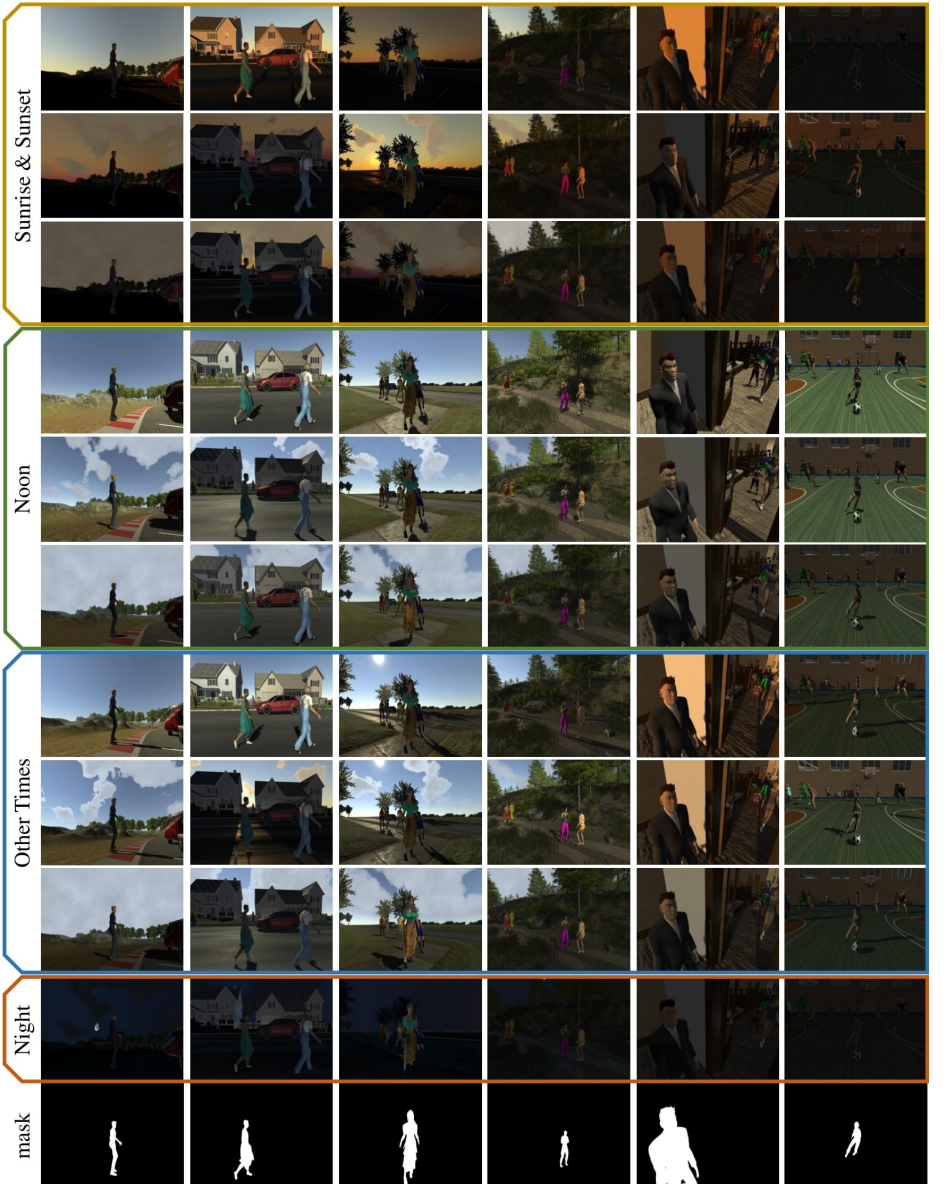



Figure 3: Example ground-truth rendered images in RdHarmony of “human” category. The left four columns are outdoor scenes (raceway, downtown, street, and forest) and the right two columns are indoor scenes (bar and stadium). Under each time of the day except “Night", from top to bottom, we show rendered images captured under Clear, Partly Cloudy, and Cloudy weather.

Figure 4: Example ground-truth rendered images of extended 5 object categories. From left to right, we show images from “bottle”, “cake”, “motorcycle”, “cow”, “couch” category. Under each time of the day except “Night", we show rendered images captured under three representative weather.

images and ground-truth rendered images. Totally, our RdHarmony has 180,000 training pairs from 6 novel categories.

### 1.4 Rich Diversity

To ensure the diversity of RdHarmony, we have put efforts on the rendered images in several ways. First, as mentioned in Section 1.2, we collect 30 3D scenes from Unity Asset Store and CG websites, including outdoor scenes (*e.g.*, raceway, downtown, street, forest) and indoor scenes (*e.g.*, bar, stadium, gym). Second, to keep in line with the large intra-category variance of "human" category, we leverage MakeHuman to create 1,500 distinct 3D human characters with controllable attributes. For 5 object categories ("bottle", "cake", "motorcycle", "cow", "couch"), we collect 3D models from Unity Asset Store and public websites. Third, for each 2D scene, we sample 10 ground-truth rendered images with 10 different capture conditions (*i.e.*, styles), including the the night style as well as styles of Clear/Partly Cloudy/Cloudy weather at sunrise&sunset/noon/other-times. *Note that each style is not a discrete style, but covers a range of illumination intensity and direction.* In Figure 3, we select some indoor/outdoor 2D scenes, and exhibit all 10 ground-truth rendered images on "human" category for each 2D scene. Besides, Figure 4 shows 10 ground-truth rendered images for each 2D scene and one 2D scene for each object category. These can provide an intuitive perspective for the diversity of our contributed rendered image dataset RdHarmony.

With a group of 10 images from the same scene, our ground-truth rendered images are quite similar to day2night [10], *i.e.*, the same scene captured under different capture conditions. Therefore, we naturally adopt the same procedure as constructing Hday2night [2] when generating composite rendered images, that being said, exchanging foregrounds within the group of images. As stated in [2, 4], Hday2night could be deemed as a real composite dataset, where the composite images are much closer to real-world applications. However, collecting a set of images for exactly the same scene under different capture conditions is not trivial [2], which limits the scale of Hday2night. On the contrary, *our dataset construction simulates the ideal generation process of training pairs for image harmonization that could rarely be implemented in real scenarios*. Our RdHarmony, which contains 180,000 pairs, could supplement real dataset to a large extent.

## 2 Implementation Details

We adopt the UNet-like architecture of iDIH as backbone considering its simplicity and effectiveness. Note that we do not use auxiliary semantic information as in [7] for brevity and fair comparison with other methods. The domain-specific encoder (*resp.*, decoder) contains the first (*resp.*, last) 4 layers in the encoder (*resp.*, decoder). The other layers form the domain-invariant encoder-decoder. After a few trials, we set $\lambda_{adv}$, $\lambda_{sty}^{rd}$, and $\lambda_{sty}^{rl}$ to 0.1, 0.1, and 0.05 respectively, by observing the harmonization quality of training images. The impact of layer configuration and hyper-parameters can be found in Section 3. Following [2], we use MSE, fMSE (foreground MSE), and PSNR as evaluation metrics. The images are resized to $256 \times 256$ in training and test phases. All the experiments except in Section 8 treat "human" as the novel category.

Our network is implemented using Pytorch 1.4.0 and trained on ubuntu 16.04 LTS operation system, with 64GB memory, Intel Core i7-8700K CPU, and two GeForce GTX 1080

| # | $\mathcal{L}_G^{rd}$ | $\mathcal{L}_{in}^{rd}$ | $\mathcal{L}_{out}^{rd}$ | $\mathcal{L}_W^{rl}$ | $\mathcal{L}_{ER}^{rl}$ | fMSE ↓ | PSNR ↑ |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | 339.02 | 36.06 |
| 2 | ✓ | | | | | 303.69 | 36.45 |
| 3 | ✓ | ✓ | ✓ | | | 300.22 | 36.50 |
| 4 | ✓ | ✓ | ✓ | ✓ | | 297.99 | 36.52 |
| 5 | ✓ | ✓ | ✓ | | ✓ | 297.94 | 36.51 |
| 6 | ✓ | ✓ | ✓ | ✓ | ✓ | 296.40 | 36.60 |

| $L$ of $E^r$ | $L$ of $D^r$ | fMSE↓ | PSNR↑ |
|---|---|---|---|
| 0 | 7 | 498.82 | 34.48 |
| 3 | 3 | 327.45 | 36.20 |
| 4 | 4 | 296.40 | 36.60 |
| 5 | 5 | 328.39 | 36.19 |
| 7 | 0 | 501.67 | 34.58 |

Table 1: Ablation studies. Left sub-table: ablation studies on the losses. Right sub-table: ablation studies on the hyper-parameter $L$. We use $E^r$ (*resp.*, $D^r$) to represent both $E^{rd}$ (*resp.*, $D^{rd}$) and $E^{rl}$ (*resp.*, $D^{rl}$). "$L$ of $E^r$ (*resp.*, $D^r$)" denotes the number of unshared layers in the first (*resp.*, third) stage. All results are tested on $\mathcal{D}_{te,n}^{rl}$.

Ti GPUs. The weight of the network is initialized with values drawn from the normal distribution $\mathcal{N}(mean = 0.0, std^2 = 0.02)$.

The discriminator $D$ consists of 3 convolution layers with channel number $\{64, 32, 1\}$, $3 \times 3$ kernels and stride of 1, each of which is followed by a Leaky-ReLU except the last one. The style classifiers $P^{in}$ and $P^{out}$ are the same in the structure and shared across domains. They contain 4 convolution layers with channel number $\{64, 32, 16, 1\}$, $3 \times 3$ kernels and stride of 1, each of which is followed by a BatchNorm, a ReLU, and a max-pooling layer except the last one.

To train the network, we use Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Learning rate is initialized with $1e^{-4}$ and reduced by a factor of 10 at epochs 100 and 120. The batch size is set to 8 and the models are trained for 150 epochs.

# 3 Ablation Studies

In this section, we will validate the effectiveness of different loss terms in Section 3.1, explore the variants of network design in Section 3.2, and analyse the impact of hyper-parameters in Section 3.3.

## 3.1 Loss Design

We ablate each component of the loss function in Table 1. When we use neither adversarial loss nor style-related losses, the network only contains two three-stage generators. Though the performance is degraded, it is still better than iDIH (row 4 in Table 2 in the main text) due to the information sharing in the second stage. By adding adversarial loss $\mathcal{L}_G^{rd}$ after the first stage, the performance is boosted, which demonstrates the efficacy of adversarial loss to pull close two domains. Moreover, after adding two style losses $\mathcal{L}_{in}^{rd}$ and $\mathcal{L}_{out}^{rd}$ in the rendered image domain, the performance is further improved, which indicates the potential of available style information. We ablate the style aggregation loss $\mathcal{L}_{SA}^{rl}$ by its two terms $\mathcal{L}_W^{rl}$ and $\mathcal{L}_{ER}^{rl}$. The results demonstrate that each term is helpful and their combination achieves further improvement with mutual collaboration.

We present example test images from $\mathcal{D}_{te,n}^{rl}$ harmonized under four different loss designs, including $\mathcal{L}_{rec}^{rd} + \mathcal{L}_{rec}^{rl}$ (row 1 of left sub-table in Table 1), $\mathcal{L}_{rec}^{rd} + \mathcal{L}_{rec}^{rl} + \mathcal{L}_G^{rd}$ (row 2), $\mathcal{L}_{rec}^{rd} + \mathcal{L}_{rec}^{rl} + \mathcal{L}_G^{rd} + \mathcal{L}_{in}^{rd} + \mathcal{L}_{out}^{rd}$ (row 3), and our full method $\mathcal{L}_{rec}^{rd} + \mathcal{L}_{rec}^{rl} + \mathcal{L}_G^{rd} + \mathcal{L}_{in}^{rd} + \mathcal{L}_{out}^{rd} + \mathcal{L}_{SA}^{rl}$ (row 6). The results are shown in Figure 5. We observe that our CharmNet could produce more harmonious results that are closer to the ground-truth real images. By comparing the

Figure 5: Example results generated under different loss designs on $\mathcal{D}^{rl}_{te,n}$. From left to right, we show the input composite real image, ground-truth real image, as well as the harmonized results generated by Sim 1, Sim 2, Sim 3, and our full method CharmNet, where Sim 1 $= \mathcal{L}^{rd}_{rec} + \mathcal{L}^{rl}_{rec}$, Sim 2 $= \mathcal{L}^{rd}_{rec} + \mathcal{L}^{rl}_{rec} + \mathcal{L}^{rd}_{G}$, Sim 3 $= \mathcal{L}^{rd}_{rec} + \mathcal{L}^{rl}_{rec} + \mathcal{L}^{rd}_{G} + \mathcal{L}^{rd}_{in} + \mathcal{L}^{rd}_{out}$, and Full Version $= \mathcal{L}^{rd}_{rec} + \mathcal{L}^{rl}_{rec} + \mathcal{L}^{rd}_{G} + \mathcal{L}^{rd}_{in} + \mathcal{L}^{rd}_{out} + \mathcal{L}^{rl}_{SA}$. The foregrounds are outlined in red.

results of "Sim 1" and "Sim 2" in Figure 5, it can be seen that adversarial loss $\mathcal{L}^{rd}_{G}$ is important for aligning two domains to generate reasonable harmonized results. By comparing the results of "Sim 2", "Sim 3" and "Full Version", it can be seen that style classification losses $\mathcal{L}^{rd}_{in} + \mathcal{L}^{rd}_{out}$ in the rendered image domain and the style aggregation loss $\mathcal{L}^{rl}_{SA}$ in the real image domain are both meaningful for image harmonization, and their combination ensures the useful knowledge transfer across domains, thus leading to more satisfactory performance.

## 3.2 Network Design

We analyze the network design in terms of the number of unshared layers in the domain-specific encoding stage and domain-specific decoding stage. As mentioned in Section 4.1 in the main text, we empirically split the first (*resp.*, last) $L$ layers in encoder (*resp.*, decoder) into the first (*resp.*, third) stage. Therefore, we also investigate the impact of hyper-parameter $L$, and the results are also reported in Table 1. When we increase $L$ to 5, the performance is degraded, since it hampers the useful information sharing in the second stage. In addition, when we decrease $L$ to 3, the performance is also degraded, possibly because it brings difficulty in aligning the extracted features from two domains. We also explore the extreme cases where all the encoder or decoder layers are shared across domains. Setting $L$ of $E^r$ (*resp.*, $D^r$) to 0 means sharing the entire encoder (*resp.*, decoder). The performances in both cases are significantly dropped. This is because the input space and output space for image harmonization are the same in each domain but different across domains. Simply sharing the entire encoder or decoder will hurt the cross-domain transfer.

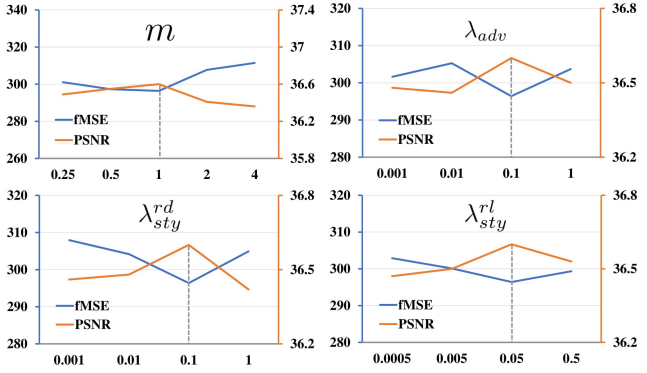



Figure 6: Impact of hyper-parameters, including the margin $m$ in Eqn. 5 and trade-off parameters $\lambda_{adv}$, $\lambda_{sty}^{rd}$, and $\lambda_{sty}^{rl}$ in Eqn. 6 in the main text. The results are tested on $\mathcal{D}_{te,n}^{rl}$ and the gray dotted line indicates the default value of each hyper-parameter.

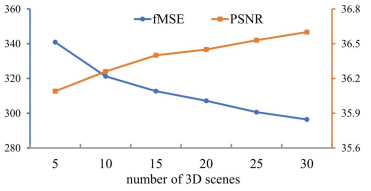
### 3.3 Hyper-parameter Analyses

We investigate the impact of four hyper-parameters: the margin $m$ in Eqn. 5 and trade-off parameters $\lambda_{adv}$, $\lambda_{sty}^{rd}$, and $\lambda_{sty}^{rl}$ in Eqn. 6 in the main paper. In Figure 6, we plot the performance on $\mathcal{D}_{te,n}^{rl}$ by varying each hyper-parameter while keeping the other hyper-parameters fixed. It can be seen that our method is robust with hyper-parameters in reasonable ranges (*i.e.*, $m$ in range $[2^{-2}, 2^{2}]$, $\lambda_{adv}$ and $\lambda_{sty}^{rd}$ in range $[10^{-3}, 1]$ and $\lambda_{sty}^{rl}$ in range $[5\times10^{-4}, 5\times10^{-1}]$).

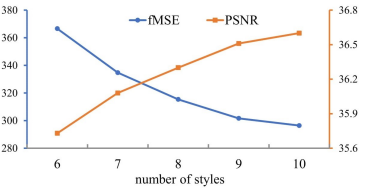
## 4 Training Data Analyses

In the dataset construction, to ensure the diversity, we collect 30 3D scenes to cover various virtual environments and define 10 representative styles to cover the majority of the day (we do not sample the whole night since there is little variation when it is totally dark) and 3 representative weather. Note that each style covers a range of illumination intensity and direction. To investigate the influence of diversity of constructed RdHarmony, we explore the performance variance by using different numbers of 3D scenes and styles. To be specific, for 3D scenes, we use the full set of $\mathcal{D}_{tr,b}^{rl}$ and a subset of $\mathcal{D}_{tr,n}^{rd}$ (*i.e.*, $\mathcal{D}_{tr,n}^{rd}(sub)$) which belongs to a subset of 30 scenes. Similarly, for styles, we use $\mathcal{D}_{tr,b}^{rl}$ and $\mathcal{D}_{tr,n}^{rd}(sub)$ which contains a subset of 10 styles. The results are shown in Figure 7(a) and Figure 7(b). From Figure 7(a), we can observe that the performance grows better when using more scenes. As the number of 3D scenes increases up to 30, we can also observe a trend in convergence. Since for each 3D scene we set 50 camera viewpoints to shot 2D scenes, an increase in 3D scene numbers leads to a much larger increase in 2D scene numbers, which makes the dataset more and more diverse. From Figure 7(b), we have a similar observation. When using fewer styles, since it could not cover both time-of-the-day and weathers well, the performance is inferior. When the number of styles increases to 10, the performance is also improved and getting stable at a superior level. Based on such observation, we use all 30 3D scenes and 10 styles in all experiments which use $\mathcal{D}_{tr,n}^{rd}$.

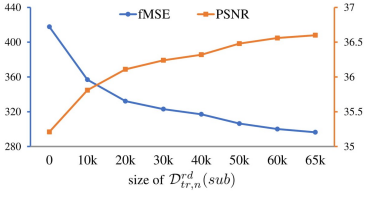
In the experiments in the main paper, we use $65k$ rendered training images with "human" category in $\mathcal{D}_{tr,n}^{rd}$. Here, we explore the performance variance when using different



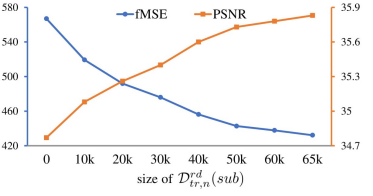

(a) Train on different numbers of scenes. Test on $\mathcal{D}^{rl}_{te,n}$.



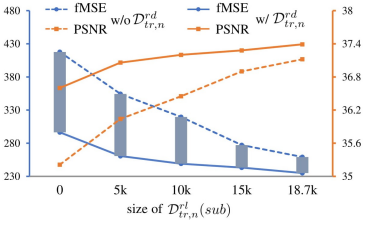

(b) Train on different numbers of styles. Test on $\mathcal{D}^{rl}_{te,n}$.



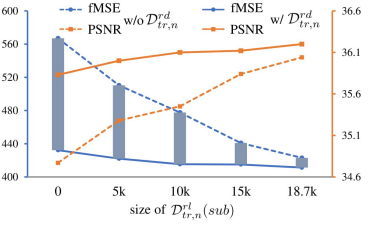

(c) Train on $\mathcal{D}^{rl}_{tr,b} + \mathcal{D}^{rd}_{tr,n}(sub)$. Test on $\mathcal{D}^{rl}_{te,n}$.




(d) Train on $\mathcal{D}^{rl}_{tr,b} + \mathcal{D}^{rd}_{tr,n}(sub)$. Test on $\mathcal{D}^{rl}_{te,b}$.




(e) Train on $\mathcal{D}^{rl}_{tr,b} + \mathcal{D}^{rl}_{tr,n}(sub)$ with and without $\mathcal{D}^{rd}_{tr,n}$. Test on $\mathcal{D}^{rl}_{te,n}$.




(f) Train on $\mathcal{D}^{rl}_{tr,b} + \mathcal{D}^{rl}_{tr,n}(sub)$ with and without $\mathcal{D}^{rd}_{tr,n}$. Test on $\mathcal{D}^{rl}_{te,b}$.

Figure 7: Top row: training with a subset of 10 styles and 30 3D scenes. Mid row: training with a subset of $\mathcal{D}^{rd}_{tr,n}$ (*i.e.*, $\mathcal{D}^{rd}_{tr,n}(sub)$) and the full set of $\mathcal{D}^{rl}_{tr,b}$. Bottom row: training with a subset of $\mathcal{D}^{rl}_{tr,n}$ (*i.e.*, $\mathcal{D}^{rl}_{tr,n}(sub)$) and the full set of $\mathcal{D}^{rl}_{tr,b}$ with and without the full set of $\mathcal{D}^{rd}_{tr,n}$.

numbers of rendered training images. Specifically, we use the full set of $\mathcal{D}^{rl}_{tr,b}$ and a subset of $\mathcal{D}^{rd}_{tr,n}$ (*i.e.*, $\mathcal{D}^{rd}_{tr,n}(sub)$) with various numbers of rendered training images to train our CharmNet, and evaluate the model on both $\mathcal{D}^{rl}_{te,n}$ and $\mathcal{D}^{rl}_{te,b}$. As shown in Figure 7(c) and Figure 7(d), the performance is improved with an increasing size of $\mathcal{D}^{rd}_{tr,n}(sub)$. Though the performance growth slows down when the size approaches $65k$, the best performance is still obtained under the full set of $\mathcal{D}^{rd}_{tr,n}$. Note that introducing more rendered training images will bring marginal improvement yet higher computational cost, so we use $65k$ rendered training images by default.

Besides, we explore another experimental setup, where real training set has inadequate number of examples from novel category. *It is very common that the real training set has a few yet insufficient examples for certain categories, in which case we can enrich the real training set with rendered images from these categories.* Specifically, we use the full set of $\mathcal{D}^{rl}_{tr,b}$ and a subset of $\mathcal{D}^{rl}_{tr,n}$ (*i.e.*, $\mathcal{D}^{rl}_{tr,n}(sub)$) with various numbers of real training images from novel category to train the iDIH [7] backbone. In addition, for each size of $\mathcal{D}^{rl}_{tr,n}(sub)$, we

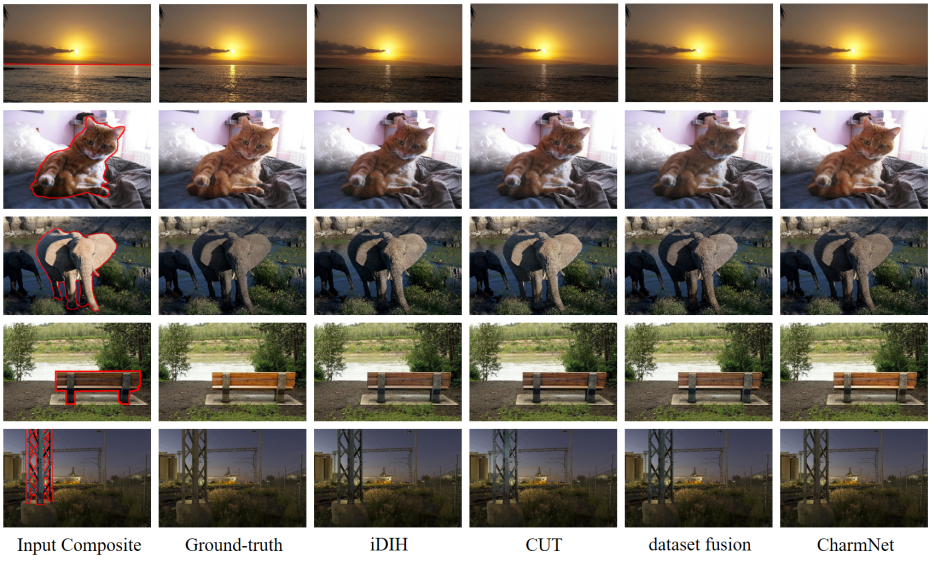



Figure 8: Example results generated by different baselines and our method on $\mathcal{D}^{rl}_{te,b}$. From left to right, we show the input composite real image, ground-truth real image, as well as the harmonized images generated by iDIH [7] backbone, CUT [6], dataset fusion, and our CharmNet. The foregrounds are outlined in red.

also augment the training set with the full set of $\mathcal{D}^{rd}_{tr,n}$ to train our CharmNet. The trained models are evaluated on both $\mathcal{D}^{rl}_{te,n}$ and $\mathcal{D}^{rl}_{te,b}$ for comparison. The results are shown in Figure 7(e) and Figure 7(f), from which we have three observations. Firstly, no matter using rendered images or not, increasing the size of $\mathcal{D}^{rl}_{tr,n}(sub)$ will bring consistent performance improvement. Secondly, with a specific size of $\mathcal{D}^{rl}_{tr,n}(sub)$ (*e.g.*, $5k$), using rendered training images $\mathcal{D}^{rd}_{tr,n}$ from novel category could boost the performance. Thirdly, when the size of $\mathcal{D}^{rl}_{tr,n}(sub)$ is small, the performance gain brought by $\mathcal{D}^{rd}_{tr,n}$ and our CharmNet is significant, which demonstrates that *auxiliary rendered images are especially helpful when the real training set is short of real images from novel categories.*

# 5 Qualitative Analyses on Base Categories

In Section 5.3 in the main text, we observe that our CharmNet not only boosts the performance on novel category, but also significantly enhances the performance on base categories. Therefore, given a composite real image from $\mathcal{D}^{rl}_{te,b}$, we also provide the harmonization results generated by iDIH [7] (row 4 in Table 2 in the main text), CUT [6], dataset fusion, and our CharmNet for comparison. As shown in Figure 8, harmonization results of our CharmNet are more plausible and closer to the ground-truth real images. It could be observed that even if the foreground is not "human", the foreground style could be well-adapted to the background style, which indicates that *the cross-domain style knowledge transfer is also useful for cross-category harmonization.* This observation also demonstrates the potential of using rendered images for image harmonization task.

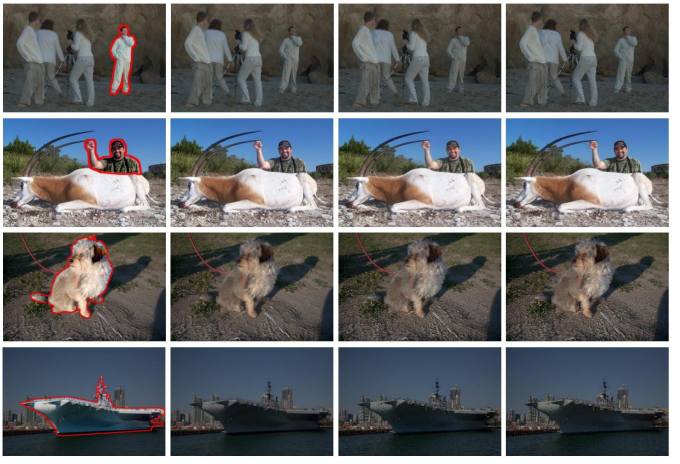



Figure 9: Example results generated by iDIH upper bound and our method. The top two rows are images from $\mathcal{D}^{rl}_{te,n}$ and the bottom two rows are images from $\mathcal{D}^{rl}_{te,b}$. From left to right, we show the input composite real image, ground-truth real image, as well as the harmonized images generated by the upper bound and our CharmNet. Foregrounds are outlined in red.

# 6 Comparison between Our CharmNet and the Upper Bound

In Table 2 in the main text, our CharmNet outperforms all the domain adaptation baselines and achieves a closer performance to the iDIH [7] trained with both $\mathcal{D}^{rl}_{tr,n}$ and $\mathcal{D}^{rl}_{tr,b}$, which serves as an upper bound. Therefore, we sample real test images from both $\mathcal{D}^{rl}_{te,n}$ and $\mathcal{D}^{rl}_{te,b}$ and show the ground-truth real images as well as the harmonized results of our CharmNet and the upper bound in Figure 9. It can be observed that *our CharmNet could generate harmonious results close to the upper bound for both novel and base categories*, which demonstrates the efficacy of our cross-domain harmonization method.

# 7 Results on Real-world Composite Images with Human Foregrounds

In practice, image harmonization is expected to tackle with real-world composite images, whose foreground is cut from one image and pasted on another background image. In such a scenario, there are no corresponding ground-truth images, so it is infeasible to evaluate model performance quantitatively. Following [2, 3, 8], we conduct user study on 46 real-world composite images with human foregrounds which are selected from 99 real-world composite images released by [8], and compare the harmonized results generated by iDIH [7] (row 4 in Table 2 in the main text), CUT [6], dataset fusion, and our CharmNet using subjective evaluation.

Specifically, for each real-world composite image, we could obtain five images $\{\mathbf{I}_i|_{i=1}^{5}\}$ including itself and four harmonized outputs generated by four above-mentioned methods.

| Method | Input composite | iDIH [7] | CUT [6] | dataset fusion | CharmNet |
|---|---|---|---|---|---|
| B-T score↑ | -1.191 | 0.250 | 0.137 | 0.326 | 0.479 |

Table 2: B-T scores of different methods and our CharmNet on 46 real-world composite images with human foregrounds.

| # | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Training data | | - | $\mathcal{D}^{rl}_{tr,b}$ | $\mathcal{D}^{rd}_{tr,n(H)}$ & $\mathcal{D}^{rl}_{tr,b}$ | $\mathcal{D}^{rd}_{tr,n}$ & $\mathcal{D}^{rl}_{tr,b}$ | $\mathcal{D}^{rl}_{tr,n}$ & $\mathcal{D}^{rl}_{tr,b}$ |
| $\mathcal{D}^{rl}_{te,n(Cake)}$ | fMSE↓ | 1293.83 | 439.33 | 465.49 | 425.63 | 391.74 |
| | PSNR↑ | 29.72 | 33.93 | 33.71 | 34.07 | 34.43 |
| $\mathcal{D}^{rl}_{te,n(Bottle)}$ | fMSE↓ | 1604.52 | 661.81 | 656.73 | 626.48 | 622.39 |
| | PSNR↑ | 32.78 | 35.69 | 35.83 | 36.01 | 36.05 |
| $\mathcal{D}^{rl}_{te,n(Motorcycle)}$ | fMSE↓ | 936.30 | 326.49 | 322.99 | 315.62 | 285.76 |
| | PSNR↑ | 32.28 | 35.91 | 35.94 | 36.07 | 36.56 |
| $\mathcal{D}^{rl}_{te,n(Cow)}$ | fMSE↓ | 921.92 | 411.46 | 401.94 | 374.70 | 382.79 |
| | PSNR↑ | 34.40 | 37.36 | 37.42 | 37.53 | 37.80 |
| $\mathcal{D}^{rl}_{te,n(Couch)}$ | fMSE↓ | 1565.55 | 575.16 | 563.49 | 531.70 | 504.99 |
| | PSNR↑ | 29.66 | 33.09 | 33.41 | 33.79 | 33.81 |

Table 3: Results of CharmNet trained on various training data and tested on $\mathcal{D}^{rl}_{te,n}$. "-" denotes metrics directly tested on composite real images. $\mathcal{D}^{rd}_{tr,n(H)}$ denotes rendered training images from "human" category. Note that column 5 using real training images $\mathcal{D}^{rl}_{tr,n}$ from novel category serves as the upper bound.

Then we can construct image pairs $(\mathbf{I}_i, \mathbf{I}_j)$ by randomly selecting two images from $\{\mathbf{I}_i|_{i=1}^{5}\}$. Based on 46 real-world composite images with human foregrounds, we could construct abundant image pairs for user study. Then we invite 13 users to participate in the study. We ask each user to see one image pair each time and pick out the more harmonious image in the pair. Finally, we collect 5980 pairwise results and employ the Bradley-Terry (B-T) model [1, 5] to obtain the overall ranking of all methods. As reported in Table 2, our CharmNet achieves the highest B-T score and once again outperforms other baselines.

Besides, we exhibit some example results of real-world composite images with human foregrounds used in our user study. We compare the real-world composite images with harmonization results generated by iDIH [7] (row 4 in Table 2 in the main text), CUT [6], dataset fusion, and our CharmNet. As shown in Figure 10, our method is capable of generating more favorable and satisfactory results than other methods.

## 8 Experiments on More Novel Categories

In this paper, we mainly take "human" as the example novel category to demonstrate the effectiveness of our CharmNet. In this section, we conduct experiments on more novel categories to show the generalization ability of our method. As mentioned in Section 1.3, we extend our RdHarmony dataset to 5 more categories ("bottle", "cake", "motorcycle", "cow", and "couch") from 5 different super-categories. Each category has 9,000 pairs of composite rendered images and ground-truth rendered images. Next, we conduct experiments by using

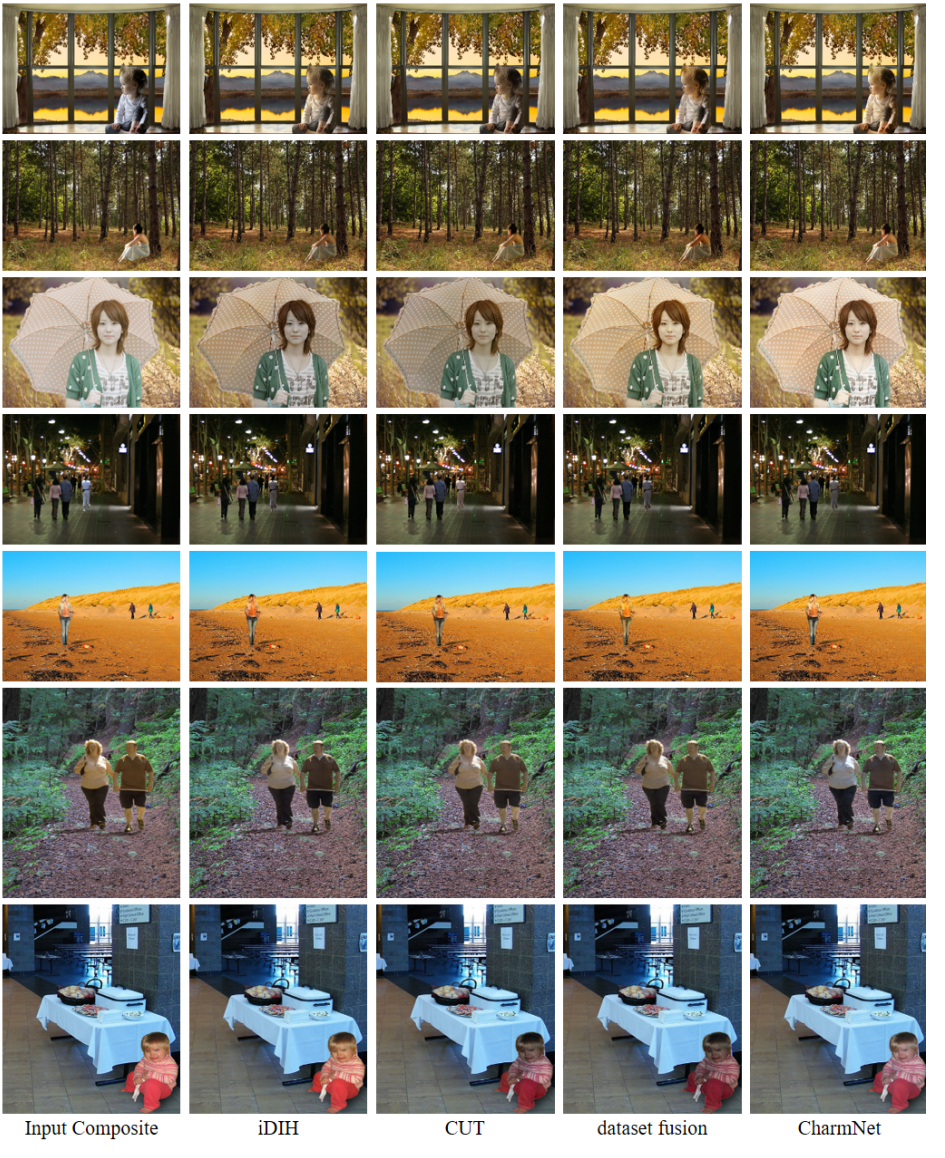



Figure 10: Example results generated by different baselines and our method on real-world composite images with human foregrounds. From left to right, we show the input real-world composite image and the harmonized results generated by iDIH [7] backbone (row 4 in Table 2 in the main paper), CUT [6], dataset fusion, and our CharmNet.

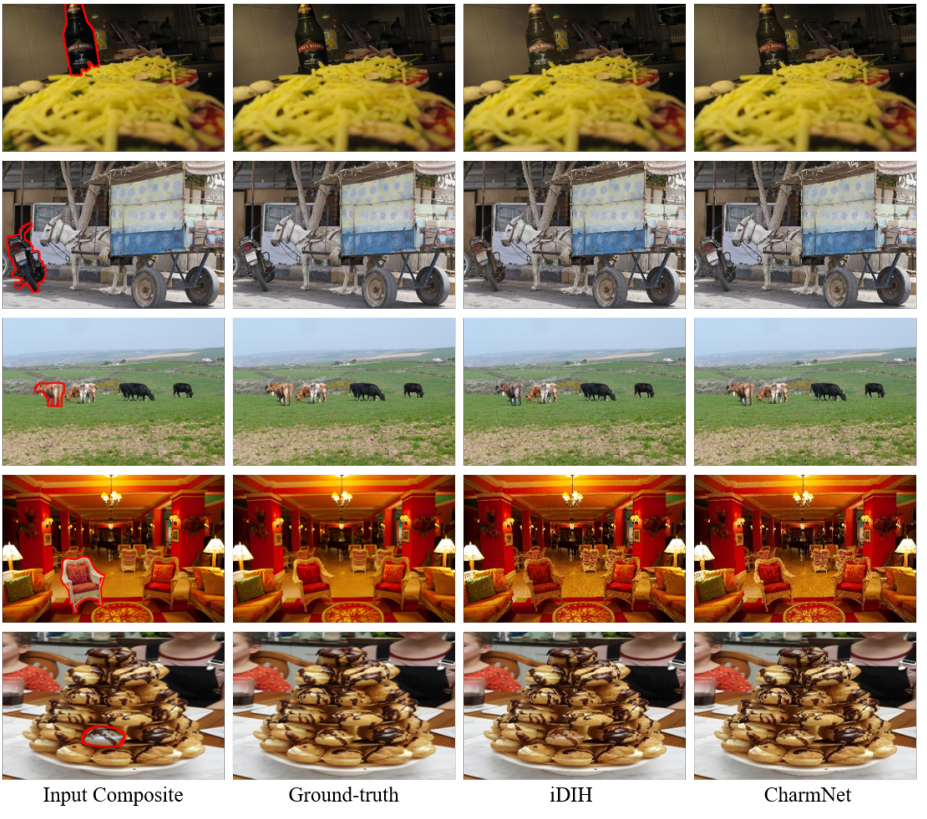



Figure 11: Example results on 5 more novel categories. From top to bottom, we show one example from "bottle", "motorcycle", "cow", "couch", and "cake" category respectively. From left to right, we show the input composite real image, ground-truth real image, as well as the harmonized images generated by iDIH [7] backbone, and our CharmNet. The foregrounds are outlined in red.

these 5 categories as example novel categories one by one and the procedure is basically the same as that with "human" novel category. By taking "bottle" as an example novel category, we treat the other categories as base categories. The training set ($\mathcal{D}^{rd}_{tr,n}$ & $\mathcal{D}^{rl}_{tr,b}$) and the test sets ($\mathcal{D}^{rl}_{te,n}$ and $\mathcal{D}^{rl}_{te,b}$) of "bottle" novel category are obtained following the same procedure as those of "human" novel category in the main text. Our training set consists of the novel training set $\mathcal{D}^{rd}_{tr,n}$ with rendered images from "bottle" category and the base training set $\mathcal{D}^{rl}_{tr,b}$ with real training images from other categories. We evaluate on the novel test set $\mathcal{D}^{rl}_{te,n}$ with real test images from "bottle" category. The results of this setting are reported in column 4 in Table 3. We compare with the results only using base training set $\mathcal{D}^{rl}_{tr,b}$ (column 2). To prove that within-category cross-domain transfer is more effective than cross-category cross-domain transfer, we also replace the novel training set $\mathcal{D}^{rd}_{tr,n}$ with equal number of rendered images from "human" category (9,000 images from RdHarmony) and report the results in column 3. Finally, we report the upper bound results using both base training set $\mathcal{D}^{rl}_{tr,b}$ and real training images from "bottle" category $\mathcal{D}^{rl}_{tr,n}$ in column 5.

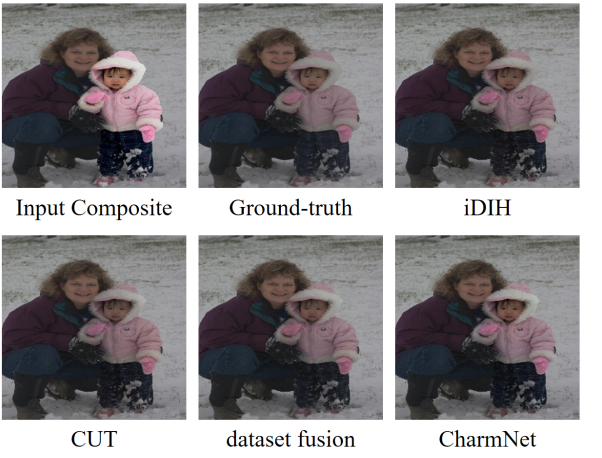



Figure 12: A sample of failure case of our CharmNet.

By comparing column 4 and column 2, we can see that using rendered training images from the target novel category can significantly boost the performance, which shows that *our method is applicable to other novel categories as well*. Based on the results in column 3, we can see that using rendered images from another category ("human") can generally improve the performance but the performance gain is much smaller and even negative (*e.g.*, "cake"), compared with the performance gain (column 4 *v.s.* column 2) using rendered images from the target novel category, which proves that *within-category cross-domain transfer is more effective than cross-category cross-domain transfer.* Despite the performance gap between column 4 and the upper bound (column 5), our method has shown great potential to bridge the gap between real images and rendered images on all 5 novel categories.

We also show the harmonization results of both iDIH backbone and our CharmNet in Figure 11, which shows that our method can generate more visually appealing results closer to the ground-truth after utilizing the rendered images from the target novel category.

# 9 Limitations

In the experiment, we observe that our CharmNet might encounter failures with the composite foreground is overexposed. For example, in Figure 12, the brim in the input composite image is overexposed. After harmonization, the result of our CharmNet may still be kind-of overexposed.

# References

[1] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.

[2] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. DoveNet: Deep image harmonization via domain verification. In *CVPR*, 2020.

[3] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Trans. Image Process.*, 29:4759–4771, 2020.

[4] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *CVPR*, 2021.

[5] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *CVPR*, 2016.

[6] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, 2020.

[7] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *WACV*, 2021.

[8] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, 2017.

[9] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[10] Hao Zhou, Torsten Sattler, and David W. Jacobs. Evaluating local features for day-night matching. In *ECCV*, 2016.