Stay Positive: Non-Negative Image Synthesis for Augmented Reality Supplemenatry Material

Katie Luo* Guandao Yang* Wenqi Xian Harald Haraldsson
Bharath Hariharan Serge Belongie
Cornell University

1. Overview

The supplementary materials in the following sections provide further implementation details, experimental details, quantitative evaluation, and qualitative results. Implementation details will be discussed in Section 2. Experiment details will be presented in Section 3. Additional quantitative results are in Section 4. Finally, qualitative results will be in Section 5, along with discussions on typical failure cases.

2. Implementation Details

2.1. Alpha (α) Values Explained

In our formal setup, α values are the percentages of light that is transmitted through the semi-transparent mirror. Under this formulation, it can be viewed as the transparency of the optical see-through device. We denote a second constant, $\beta \in [0, 1]$, which is the influence the device generated image has on the final perceptual image. For the OTS setting, we assume the amount of light that get reflected from the semi-transparent mirror is $1 - \alpha$, so the upper bound is $\beta = 1 - \alpha$. Figure 1 diagrams our basic modeling assumption.

In this model, when $\alpha = 0$, the problem falls back to video-see-through setting, as β tends to 1. Higher α values increases the problem's challenge since the amount of light that can be controlled becomes more limited. A more complete, and more complex, model would be the Bidirectional reflectance distribution function [14]. There are other optical models that may be more physically realistic that we do not explore in this work. We defer their study to future works.

*Denotes equal contribution.

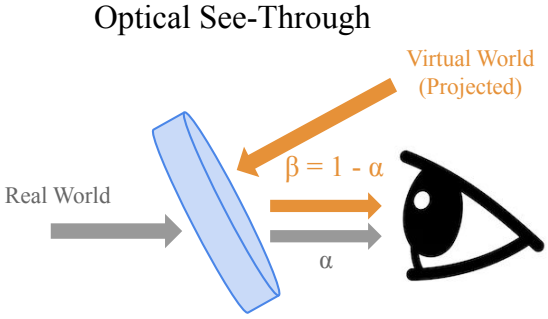


Figure 1: Our simplified optics model in the optical see-through setting.

2.2. Parameterization of Our Method

In our work, we parameterize our method using θ_1 and θ_2 as followed:

$$\begin{aligned} F_\theta(x, y) &= \theta_1 y + \theta_2, \\ O_\theta(x, y) &= \max(\min(F_\theta(x, y) - \alpha x, \beta), 0) + \alpha x. \end{aligned} \tag{1}$$

This can be viewed as trying to re-create an affine transformed version of the proposal image. Intuitively, there may be proposal images that we cannot make by purely adding light; however, we may have some hope of creating a transformed version of it due to lightness constancy [1].

We can further simplify this formulation to be a function of a single parameter, namely:

$$\begin{aligned} \theta_1 &= 1 - (\alpha x_{max} - \phi) \\ \theta_2 &= \alpha x_{max} - \phi, \end{aligned} \tag{2}$$

where x_{max} is the maximum pixel value of the input image and ϕ is the single parameter. The parameter is suggestively grouped together with the term, x_{max}. Indeed, this new parameterization has an intuitive grounding: it can be thought of as scaling our proposal image's dynamic range to be between the input image's maximum brightness αx_{max} and 1, but allowing the lower bound to *bleed into* the input image's dynamic range by an offset ϕ amount. If $\phi = 0$, we

are guaranteed our proposal image has pixel values that are greater than the input, which is desirable for non-negative image generation. However, by having the offset, ϕ, we allow the method to learn a proposal image that may not be entirely physically feasible, but increases its dynamic range. This trade-off allows for less over-exposed images.

An additional benefit is we reduce the number of parameters down to a single parameter per-image. This is desirable both for efficiency of training and explainability. In Figure 5 from the main paper, our x-axis plots loss as a function of offset ϕ variation. A natural extension is to consider the group that we are doing this transformation onto. Three that we propose are: over the entire image, over each channel, and over clusters of brightness and locality, and we report some of these results on the Map→Satellite dataset in Table 1. Note that we do not claim the superiority of one grouping strategy over another; we suggest that practitioners select the one that works best for their task.

2.3. A Clarification on Notations

In our work, we formally define two parts to our loss function, $\mathcal{L}_{const}$ and $\mathcal{L}_{sim}$,

$$\mathcal{L}_{sim}(a, b) = \|N(a) - N(b)\|^2 \quad (3)$$

$$\mathcal{L}_{const}(r, a, b) = \gamma \sum_{i,j} |\max(\min(r_{i,j}, b), a) - r_{i,j}|, \quad (4)$$

where N is a normalization function, $\gamma > 0$ is a hyperparameter controls the trade-off between perceptual similarity and fulfilling the residual constraint. In the paper we sometimes refer to $\mathcal{L}_{const}$ as $\mathcal{L}_{soft}$, as this term can be viewed as a soft constraint loss on the residual.

Furthermore, in our problem definitions, we overload the subscript notation and occasionally refer to $F(a)_{i,j}$ as $F(a_{i,j})$. This occurs in Equation 1 and Equation 3 of the main paper, and we will fix it for the final version.

Offset Group	FID (↓)	KID (↓, $\times 10^2$)	LPIPS (↓)
Channel	139.14	14.11	0.3751
Global	126.24	12.93	**0.3562**
Clustering	**125.75**	**12.76**	0.3588

Table 1: Metrics for different offset groupings reported on the Map→Satellite dataset. We do not observe a method that clearly trumps the rest.

3. Experimental Setup

3.1. Baseline Details

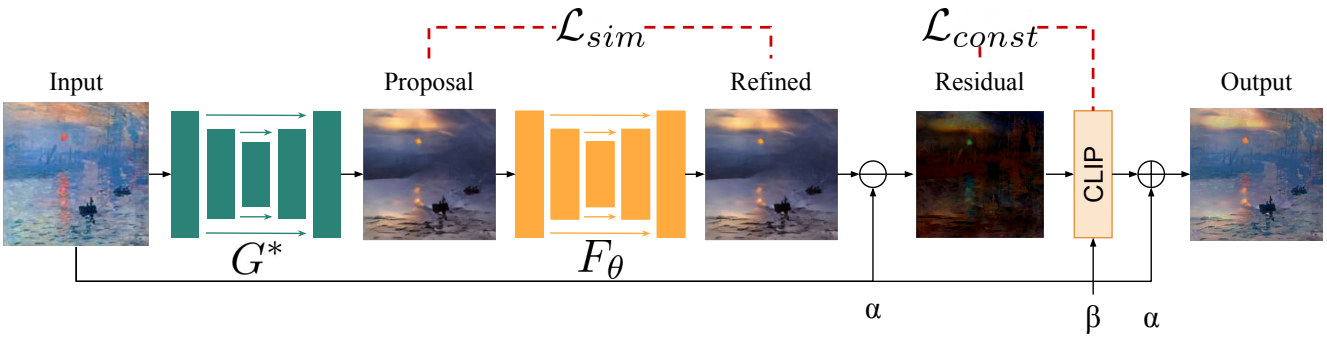
The first baseline we consider, From Scratch model (Figure 2c), is derived from the idea that the GAN discriminator loss can capture perceptual and semantic accuracy. We used the WGAN-GP [6] model to generate a residual pattern and discriminate on the optically combined output between the input and our residual. We found this to be finicky to train, even with all the bells and whistles, and tend to get trapped in sub-optimal minimas of either outputting nothing or a maxed-out residual. This is due in part to the fact that it does not have a way to elegantly handle images that do not have a non-negative residual. Visually, the From Scratch model produces blurry residuals that generally have the color scheme correct, but nothing else (Figure 5).

The second baseline, Heuristic method (Figure 2b), follows the intuition that we can simply clip the difference between the desired output and our input to be physically realizable. This was indeed a strong baseline for images that have a brightly lit output and a relatively dimmer input, but suffers from severe ghosting artifacts when the inverse is true.

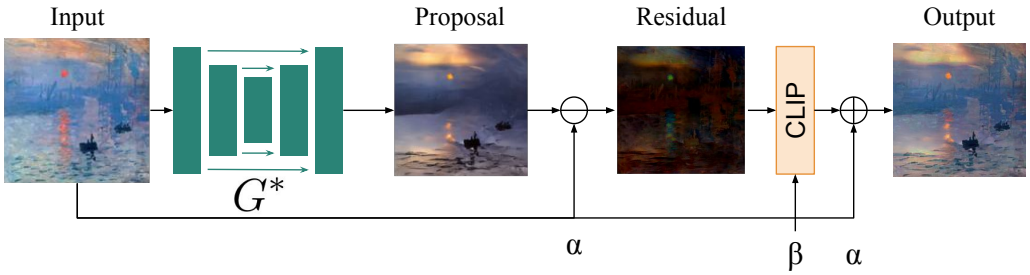
The last baseline we compare to, Finetuning model (Figure 2a), was inspired by the observation that the difference between the desired output image and the input image in the heuristic method is approximately what we want. This model thus attempts to *fine-tune* this difference into a residual that obeys the non-negativity constraint. Specifically, we train a neural network to use a pre-trained, state-of-the-art image generation network and fine-tune it such that the residual, the computed difference between the output and the input, is physically realizable. An additional loss ensures the network does not veer off from the original output. However, we experience that the finetuned model leads to severe overexposure as it gets trapped in the local optima of outputting a maxed-out residual. Tuning the parameters to optimality only leads to visually "fuzzy" outputs that deteriorate the pre-trained model's performance (Figure 5).

3.2. Implementation Details

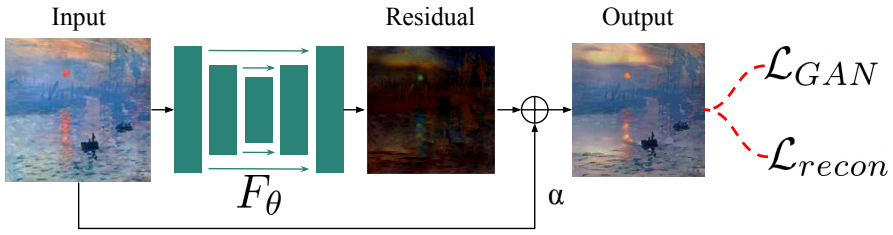
For our method, we explored different parameterization, including pure affine transformations with two parameters and single offset parameter for the entire image. We also explored grouping the offset by channel or by clustering on brightness and locality. For the optimization process, we used batch operations on batch-size of 128, and was able to compute the final results efficiently. We optimized each image to convergence, taking 5000 steps for each image; however, we observe that convergence usually occurs much quicker, usually by a one-thousand steps. The optimizer we used was Adam [12] with the default parameters and learning rate of 1×10^{-3}. During training, we also applied dropout during optimization on both the constraint loss $\mathcal{L}_{const}$ and the reconstruction loss $\mathcal{L}_{sim}$, which we found to help slightly speed up training. The reader is strongly advised to look at the code demo that is provided with this document.



(a) Finetune Baseline Model.



(b) Heuristic Baseline Method.



(c) From Scratch Baseline Model

Figure 2: Illustration of the three baselines we compare to.

3.3. Additional Metrics

Following prior works [5, 3, 10], we report two additional metrics: SSIM [18] and LPIPS [19]. SSIM measures structual similarity by comparing local patterns of pixel intensities that have been normalized for luminance and contrast. LPIPS uses deep features instead, and computes perceptual similarity by comparing the features of a VGG network. In order to simulate human visual white balancing, figures reported are normalized to be between a fixed range. To show that this carries on to real world situations, we designed a User Study with Amazon Mechanical Turk. Results are reported in Section 4.

3.4. Ablations

In Table 3 in the main paper, we do an ablation study on the Map→Satellite dataset to go over normalization, loss function importance, and our method's transformation group.

We first study the effect of the normalization module that we use,

$$N(x) = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{5}$$

where x_{min} and x_{max} are the smallest and largest pixel values in image x, respectively. Specifically, we wish to verify the claim that using the loss improves visual quality. But removing the normalization module of our method, we observe a decrease in performance, with all metrics performing worse. Visually, the output contains ghosting artifacts due to its attempt to minimize the vanilla L_2 distance.

We also study the effect of each part of the loss function on performance, $\mathcal{L}_{const}$ and $\mathcal{L}_{sim}$. Training only on $\mathcal{L}_{sim}$ yields a slightly better result as the construction of $\mathcal{L}_{sim}$ penalizes out-of-bound pixels of the target image. Training on only $\mathcal{L}_{const}$ performs poorly, as the metric we really care

about is the visual quality. The target contain very few out-of-bound pixels, but looks "overexposed" and has very little dynamic contrast. However, using both losses yields the best visual qualities, as shown in the table.

Finally, we explored results on producing images in an unconstrained fashion in row 4 of Table 3 in the main paper. For this method, we do affine transformations at a per-pixel level, allowing the image to optimize with the most parameters. Specifically, the number of parameters is equal to the size of the image. Surprisingly, this does worse than producing images under a constrained, fewer parameter setting. This is due to the fact that, by not adding constraints the original structure of the image is lost. The numbers reflect this, as all metrics except for PSNR are worse; we note that PSNR has a tendency to measure high pixel correlation as being better.

3.5. Other Applications

In this section, we will provide implementation details for how we extended our model to other applications, and precisely how we generate Figure 7 of the main paper. For Figure 7 of the paper, we extended our results to four applications. Following are the application, their corresponding papers, and the link to the implementation we used:

1. High resolution image-to-image translation [2]. We use the official implementation of this paper to generate input and proposal target image: https://github.com/saic-mdal/HiDT. We use $\alpha = 0.5$ and optimize two parameters per channel.

2. Multi-modality image-to-image translation MUNIT [8]. We use the newest released of MNUIT (https://github.com/NVlabs/imaginaire/tree/master/projects/munit) and the AFHQ dataset in StarGAN [4]. For this task we use $\alpha = 0.5$ and optimizing two global parameters.

3. Style transfer, style to photo, and sketch to photo. We use the results form AdaIn [7] for these three tasks. We use AdaIn's offical released (https://github.com/xunhuang1995/AdaIN-style) to create pairs of input and target image for our optimization procedure. For both style transfer and sketch to photo tasks, we use $\alpha = 0.7$ and optimize two parameters per channels. As for style to photo task, we use $\alpha = 0.5$.

4. Face attribute editing. For this paper, we use InterfaceGAN [16]'s official release (https://github.com/genforce/interfacegan). We use the code to generate faces using ProgressiveGAN [11]. Then we select a pair of faces, one smiled while the other didn't, as the proposal and target for our optimization procedure. For this application we use $\alpha = 0.7$ and optimizing two parameters per channel.

The baseline we shown are the Heuristic baseline. Both the baseline output and the output of our model went through basic white-balancing and contrast enhancement before showing in paper to better model perceptual comparison of human eyes.

4. Quantitative

We provide additional quantitative results to further compare our model with baselines. Specifically, we compare with baselines using different metrics (Section 4.1), using a variety of datasets (Section 4.2), and using different α value (Section 4.3). We also conduct a user study using Amazon Mechanical Turks in Section 4.4.

4.1. Additional comparison to baselines

In this section, we compute two additional metrics, SSIM [18] and LIPIS [19], on results reported in Table 1 of the main paper. Both baselines, model outputs, and the datasets are kept the same. The results are shown in Table 2. We can see that our model out-performs all baselines in LIPIS, and out-performs almost all baselines in SSIM.

4.2. Comparison in a variety of datasets

We extend the additional quantitative evaluation to datasets in CycleGAN [20], which are partially reported in Table 2 of the main paper. We present results for all datasets where pretrained model was provided by the CycleGAN codebase with SSIM and LIPIS in Table 3. The results show that our model is able to outperform the strongest baselines (i.e. Heuristic) in most tasks.

4.3. Different α's

In this section, we extending the quantitative evaluation of Figure 8 in our paper. Precisely, we extended the evaluation using two more metrics: SSIM and LIPIS. We also presented the results for three more datasets from Pix2Pix [9]: Satellite→Map, Night→Day, and Day→Night. The results are shown in Figure 4. In most metrics and datasets, our method out-performs the baseline in almost all α values. The performance of our model is very close to that of the baseline when α is very low or when α is very high, since both models works well in former cases and both models breaks in later cases. This is reasonable because when α is small, the generation task is almost the same as the unconstrained problem. When α is to large, since too much environment light has let through, there might not be feasible solution even after taking human perceptual quirks into consideration. Our model out-performs the baseline more when generating photo-realistic images, as indicated in Map→Satellite dataset comparing to the other direction.

Method	Satellite → Map [9]		Map → Satellite [9]		Day → Night [13]	
	SSIM(↑, $\times 10^2$)	LIPIS(↓)	SSIM(↑, $\times 10^2$)	LIPIS(↓)	SSIM(↑, $\times 10^2$)	LIPIS(↓)
Heuristic	33.82	0.563	35.21	0.486	56.42	0.363
From scratch	32.67	0.662	13.55	0.761	38.31	0.615
Finetuning	**42.56**	0.584	17.99	0.695	40.20	0.626
Ours	37.86	**0.534**	**49.58**	**0.359**	**61.38**	**0.343**

Table 2: Comparison with baselines using SSIM [17] and LIPIS [19] in three datasets.

Domain		SSIM (↑, $\times 10^2$)		LIPIS (↓, $\times 10^2$)	
Input	Target	Ours	Heuristic	Ours	Heuristic
Horse	Zebra	**13.70**	13.31	**75.16**	75.32
Zebra	Horse	**17.95**	16.59	74.26	**74.40**
Summer	Winter	**16.38**	16.06	**76.09**	76.15
Winter	Summer	**18.06**	16.61	**74.82**	75.01
Monet	Photo	**20.20**	18.30	**78.60**	79.19
Photo	Monet	18.00	**18.41**	**75.95**	76.19
	VonGogh	**11.09**	10.84	**74.40**	74.47
	Cezanne	**16.99**	15.89	69.92	**69.68**
	Uyeoko	17.09	**18.01**	**75.55**	75.98

Table 3: Comparison with heuristic baseline on a variety of datasets from CycleGAN [20].

The performance gap is larger when the source domain is brighter than the target domain (e.g. Day→Night).

4.4. User Study Details

While a variety of metrics have confirmed the efficacy of our proposed method, these metrics were designed to only approximate what a real user's preference is. In order to get a more accurate assessment of whether users would prefer the output of our method to the baselines, we conducted a user study using Amazon Mechanical Turks. Specifically, we random sampled 50 images from the Map→Satellite dataset and gathered outputs from our method as well as three baselines (From scratch, Finetune, and Heuristic). For each selected input image, we will presented the four output images generated by four methods in random order in a row, and asked the workers four questions:

1. Which image looks most like a satellite image?
2. Which image looks the most photo-realistic?
3. Which image is the most detailed?
4. Which image looks most like a satellite map?

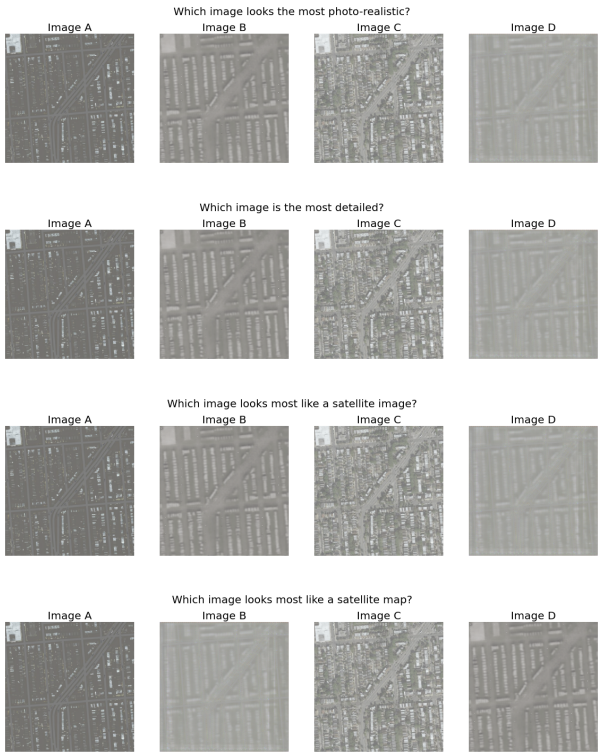


Figure 3: Examples of images for AMT workers. The workers will be presented with these four questions and asked to answer them. The last two questions were designed as a sanity check for whether the worker paid attention.

Please refer to Figure 3 for an example of images and questions presented to the workers.

To verify whether a user's response is consistent, we check whether they provide the same answer for the first and the last question. These two questions are very similar but with different order of the images. We expect a user who paid attention to the task would provide the same answer to those two question. We throw away all responses whose answer for the first and the last question aren't consistent and compute user preference using the rest data. We

receive 196 responses in total for 50 images, and rejected 73 of them (i.e. 62.76% of consistent data).

5. Qualitative

In this section, we will present additional qualitative results, comparing the proposed method with baselines on a variety of datasets and tasks. A detailed comparison between our methods and all three baselines are presented in three Pix2Pix datasets [9] in Figure 5 (Map↔Satellite and Day→Night). We ignored the Night→Day direction since such direction is usually easy unless α is so high that the task becomes infeasible. For unaligned datasets in CycleGAN [20], we show results from Zebra↔Horses in Figure 6, Winter↔Summer in Figure 8, and photo stylization in Figure 7. We used pretrained model available in https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix to generate our proposal images for those tasks. Interestingly, we found the performance is usually bottlenecked by the quality of the pretrained model in these unaligned tasks. To alleviate such issue and show that our model is capabe of combining with existing state-of-the-arts, we show additional results using most recent methods in Figure 9. One can see that our method can leverage stronger image proposal model to reach better performance.

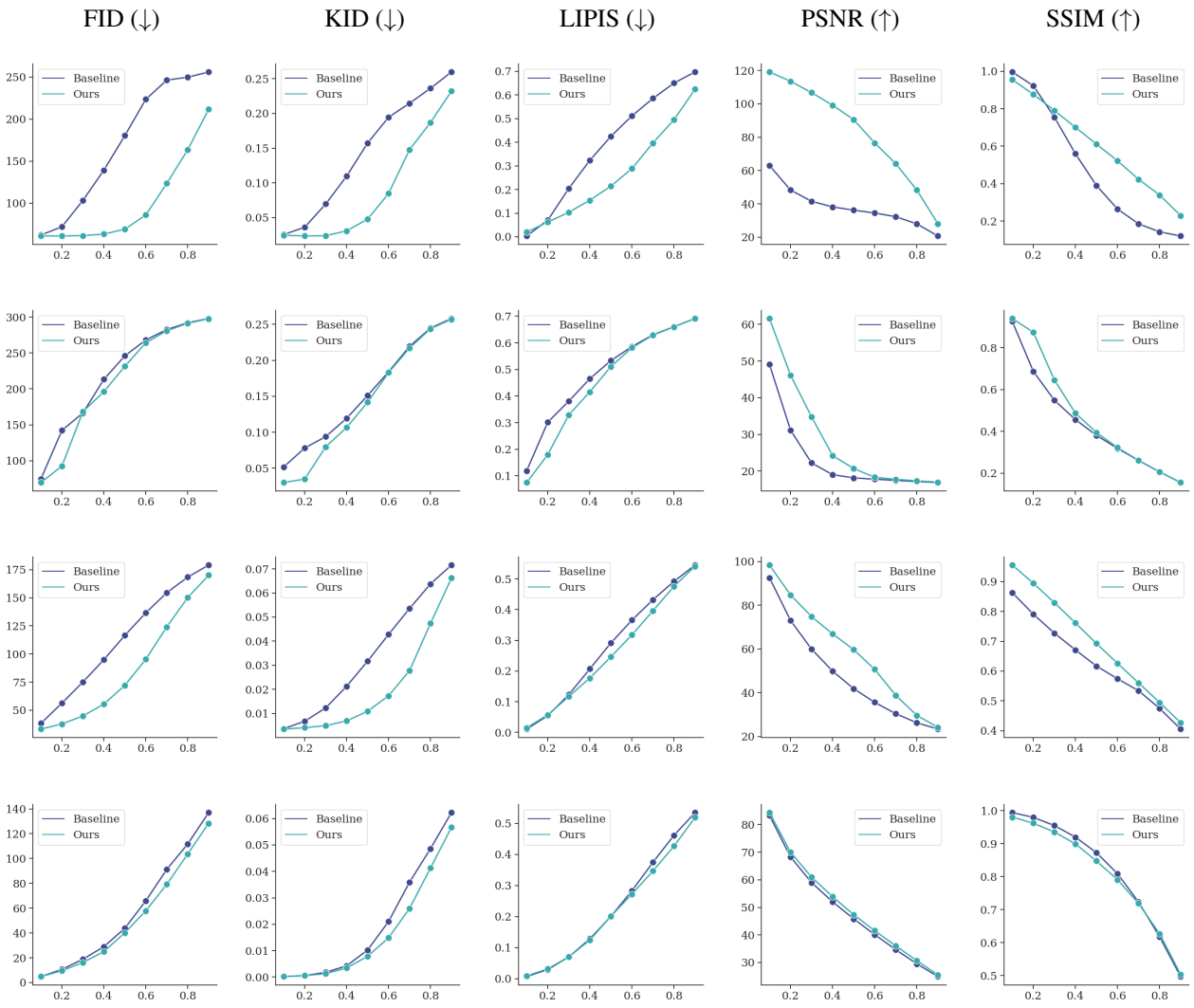


Figure 4: Metric plots over alphas. Left to right: FID, KID, LPIPS, PSNR, SSIM. Top to bottom: map→satellite, satellite→map, day→night, night→day. Our model out-performs the strongest baseline (i.e. Heuristic) across a wide range of different α value. Note that for small α, both our method and the baseline works well; while for large α, both our method and the baseline breaks done. Our method out-performs the baseline more when asking to generate photo-realistic image (i.e. map→satellite), and when the source domain is brighter than the target one (i.e. day→night).

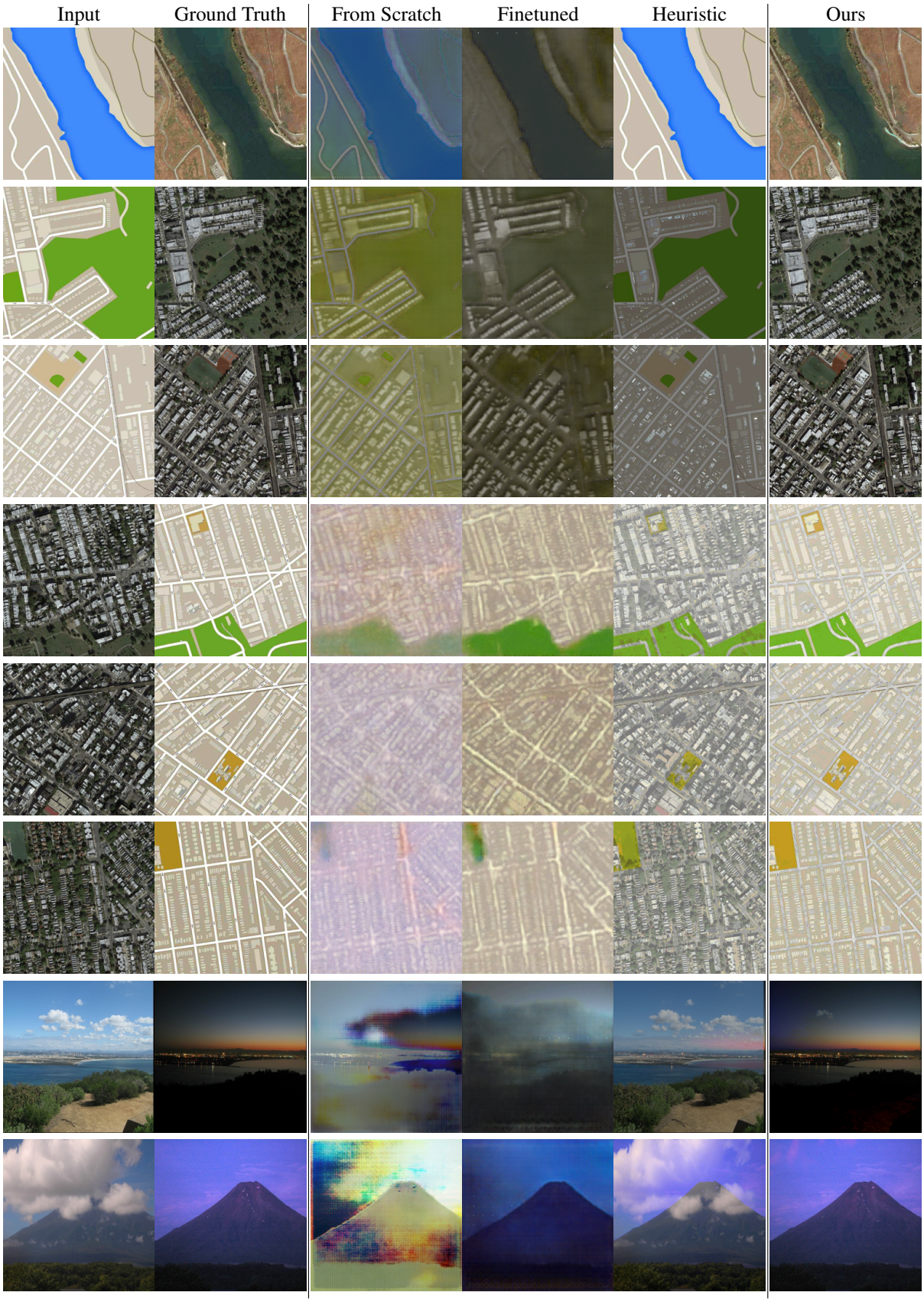


Figure 5: Visual comparison of our method on the Map↔Satellite and Day→Night dataset [9].

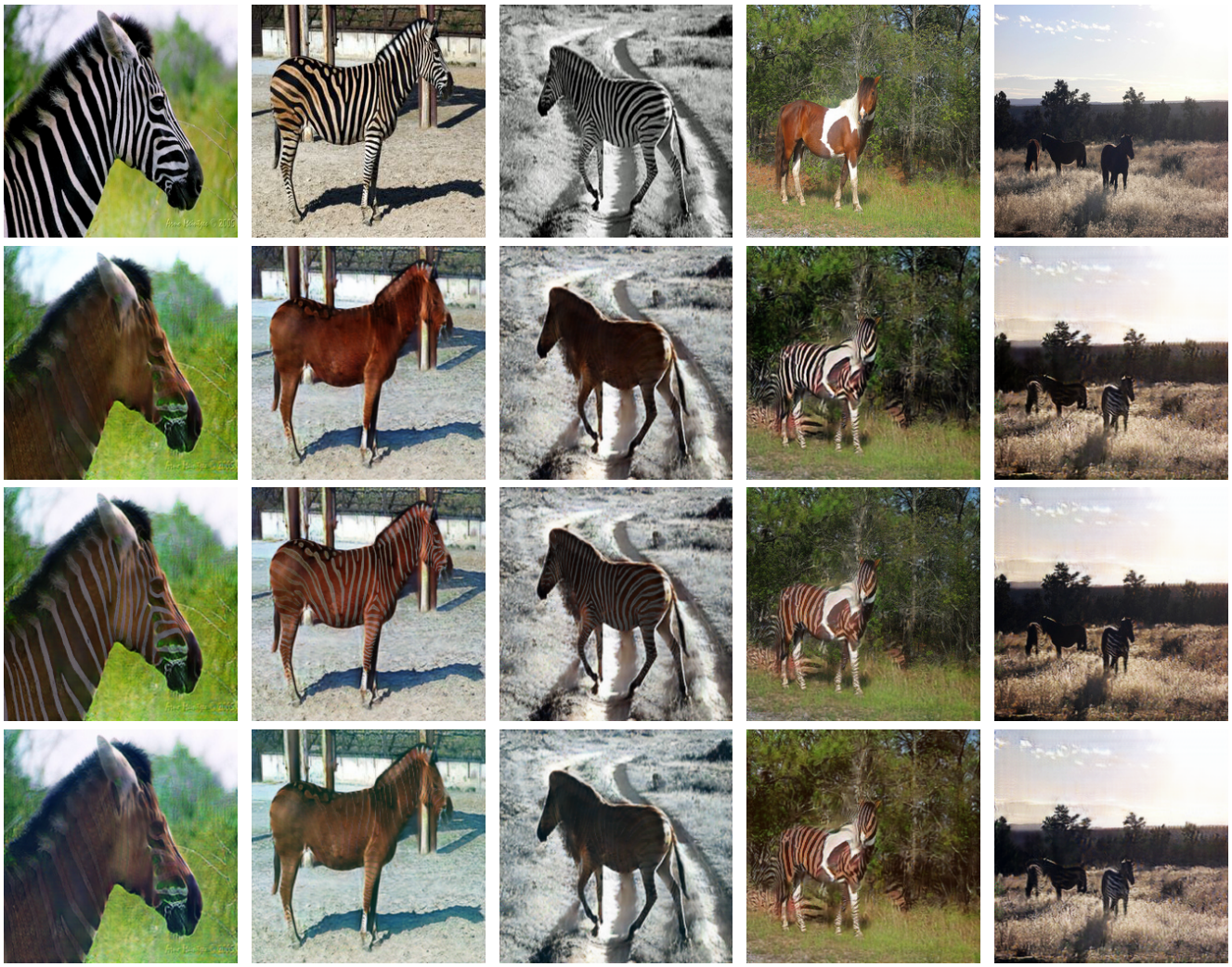


Figure 6: We compare our results on the Zebra↔Horse dataset [20]. Our method is better able to remove the white stripes on the zebras, which is difficult with a non-negative constraint. Top to bottom: input, proposal, heuristic baseline, our method.

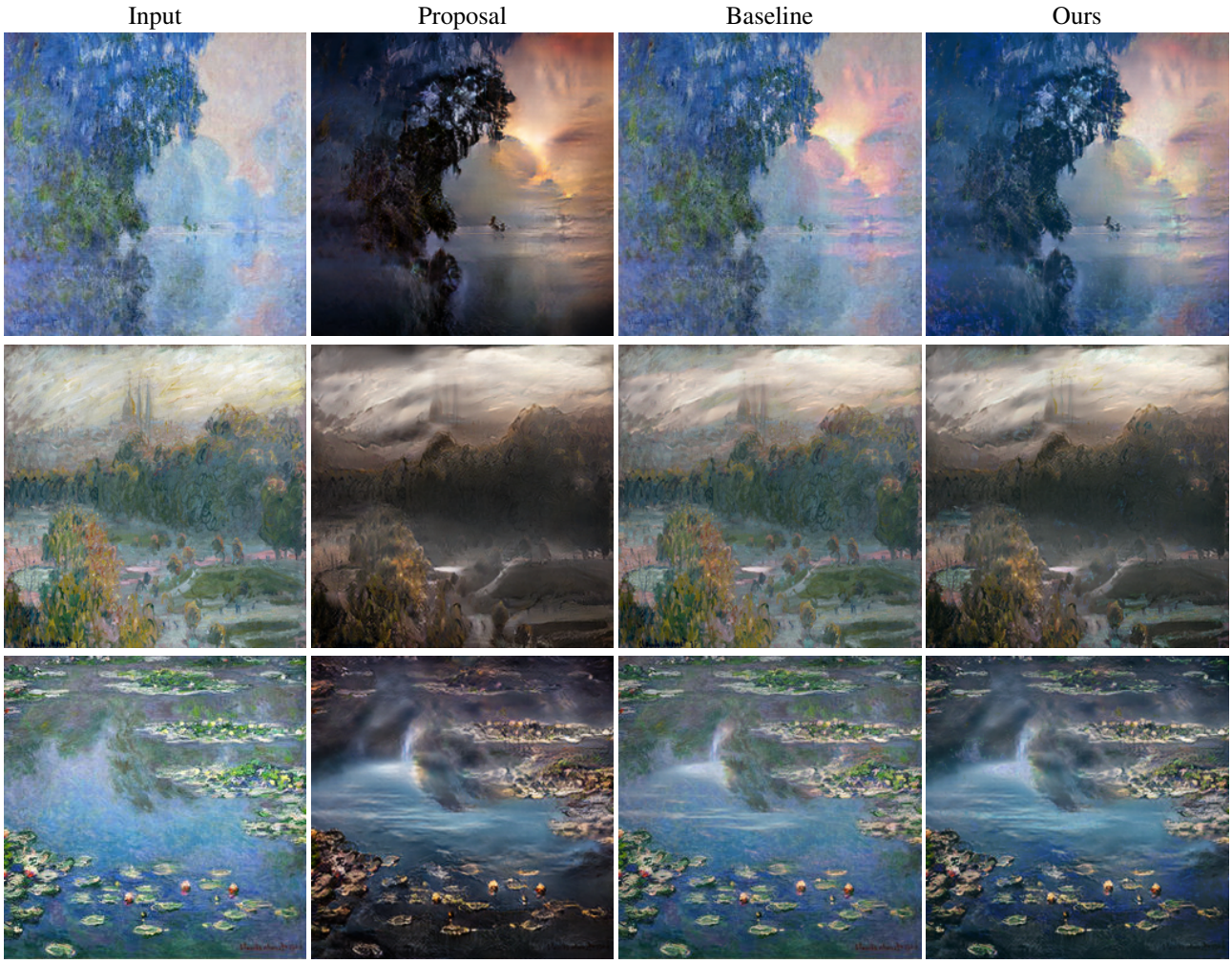


Figure 7: Visualizations of our method on Monet→Photo task [20]. Compared to naively clipping, we do not lose most of the details in our method.

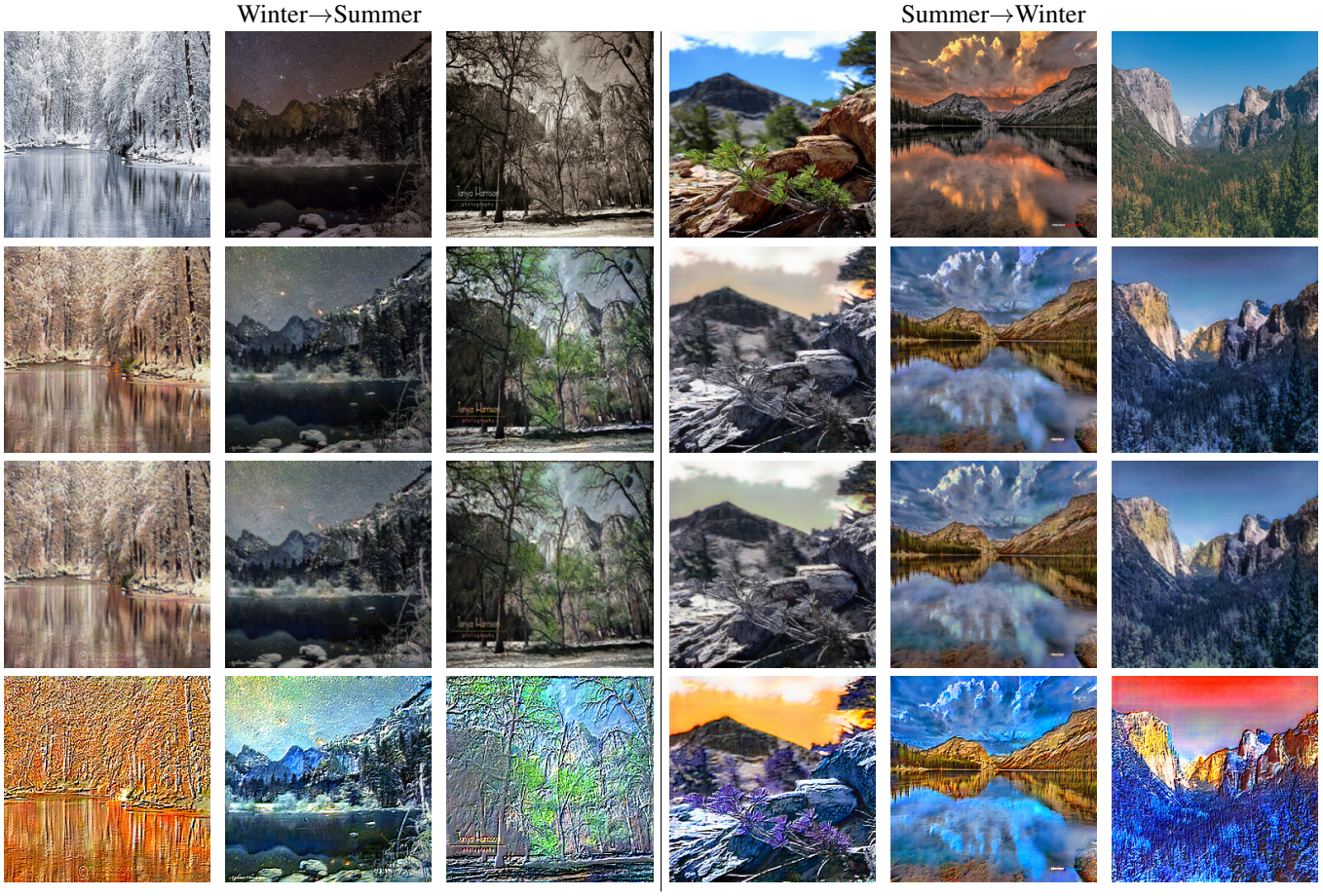


Figure 8: Visualisations on the Summer↔Winter dataset [20], along with the non-negative residual image. Our method is able to match the proposal image by generating a residual that harmoniously combines into the desired image. Top to bottom: input, proposal, our method, its residual.

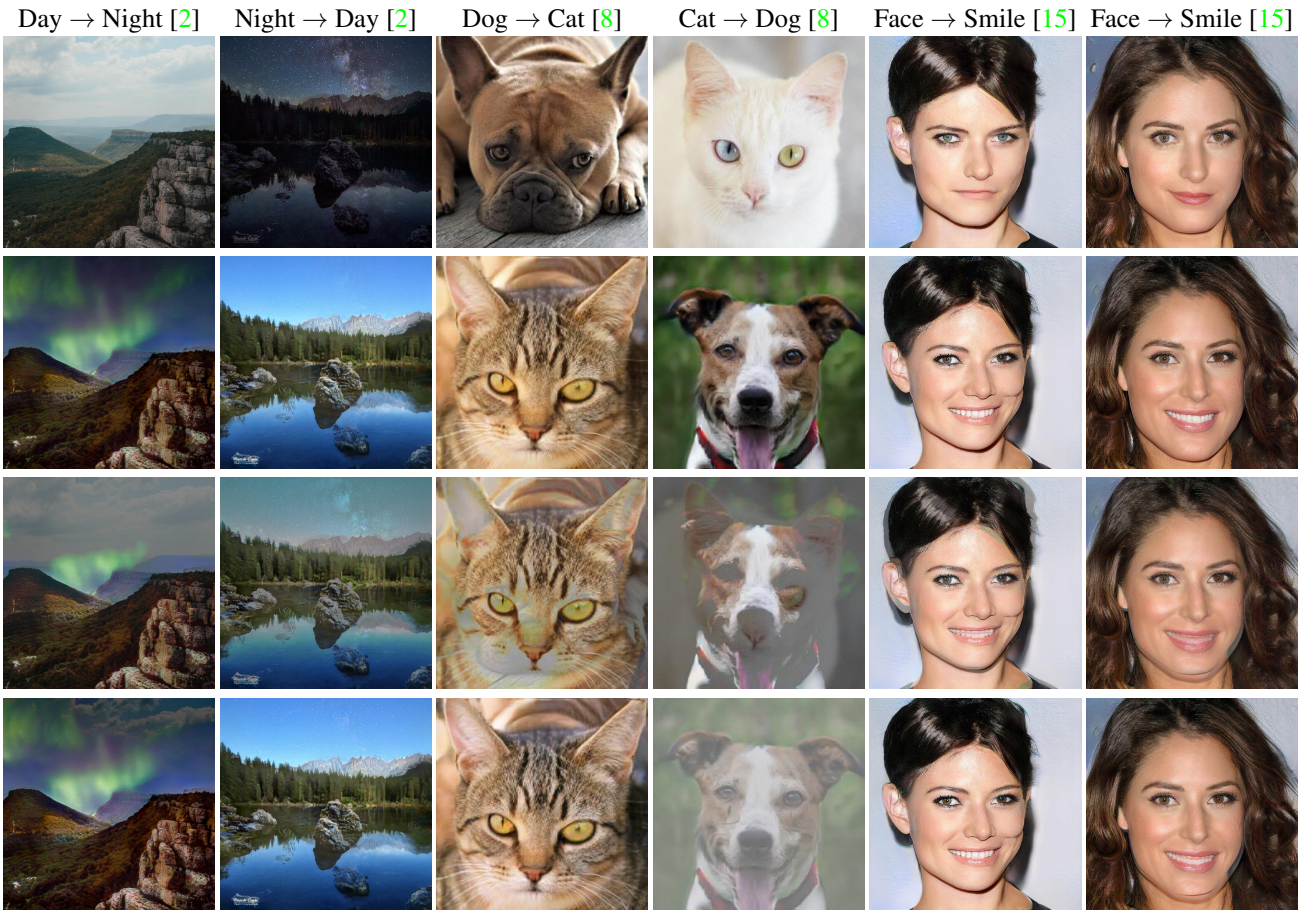


Figure 9: Additional results in different applications. From top to bottom: input, proposal, heuristic baseline, our method.

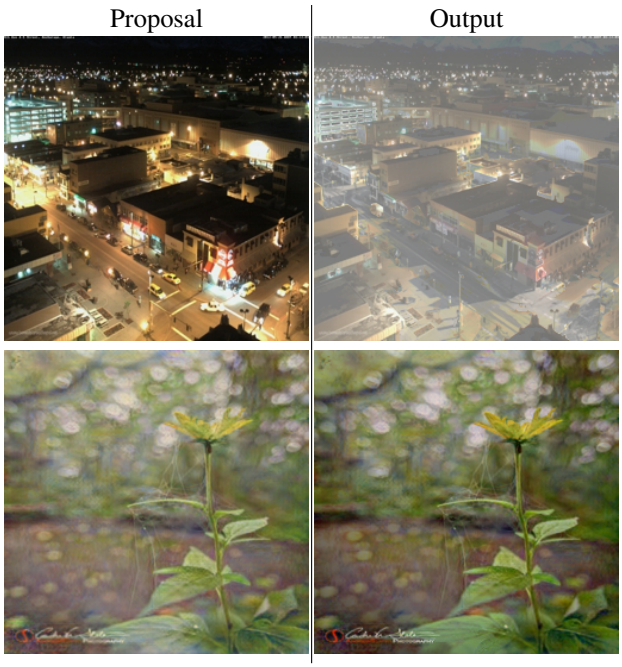


Figure 10: Failure modes of our method. It cannot deal with images with a large amount of contrast, and will look over exposed and display ghosting artifacts (top). Our method cannot recover if the image proposal is bad to begin with (bottom).

6. Failure cases

Our method works very well due to being able to leverage lightness constancy in human visual systems. In addition, by being a framework flexible enough to encompass existing image-to-image translation models.

However, the former necessarily requires that our method decrease the dynamic range in its outputs. In Figure 10, we see in the first row a night image proposal, and the corresponding output from our model. Because there are no limits on the image proposal, our method may encounter images with very large dynamic ranges. In such cases, we see that the output is riddled with ghosting artifacts, as well as looking very over-exposed in an attempt to compensate for the dark regions, but the bright portions in the proposal image prevents it from being able to naively over-saturate it. Similarly, our method has no hope to generate good images when the inputs are drastically brighter than the proposal image.

Furthermore, our framework *must* leverage image proposals. If there is no easily accessible aligned photo, then our method must rely on another image-to-image synthesis model. Thus, it cannot perform better than the synthesis model's proposals. In the second row in Figure 10, we are attempting to run a style transfer task from the photo of a plant to a Monet painting. However, the image proposal network is unable to generate anything meaningful, and despite the output of our method being close to the proposal image, it is not what we desired. In general, such limitations hold for our framework as image synthesis is not a solved problem.

References

[1] Edward H Adelson. 24 lightness perception and lightness illusions. 2000. 1

[2] Ivan Anokhin, Pavel Solovev, Denis Korzhenkov, Alexey Kharlamov, Taras Khakhulin, Alexey Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin. High-resolution daytime translation without domain labels. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4, 12

[3] Y. Blau, Roey Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor. 2018 pirm challenge on perceptual image super-resolution. *ArXiv*, abs/1809.07517, 2018. 3

[4] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 4

[5] K. Ding, K. Ma, Shiqi Wang, and Eero P. Simoncelli. A comparative study of image quality assessment models through perceptual optimization. 2020. 3

[6] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. 2

[7] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 4

[8] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *The European Conference on Computer Vision (ECCV)*, September 2018. 4, 12

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 4, 5, 6, 8

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 4

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[13] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (proceedings of SIGGRAPH)*, 33(4), 2014. 5

[14] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. A data-driven reflectance model. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH '03, page 759–769, New York, NY, USA, 2003. Association for Computing Machinery. 1

[15] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 12

[16] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *TPAMI*, 2020. 4

[17] Zhou Wang, A. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. 5

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3, 4

[19] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3, 4, 5

[20] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei 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. 4, 5, 6, 9, 10, 11