High-Resolution Daytime Translation without Domain Labels

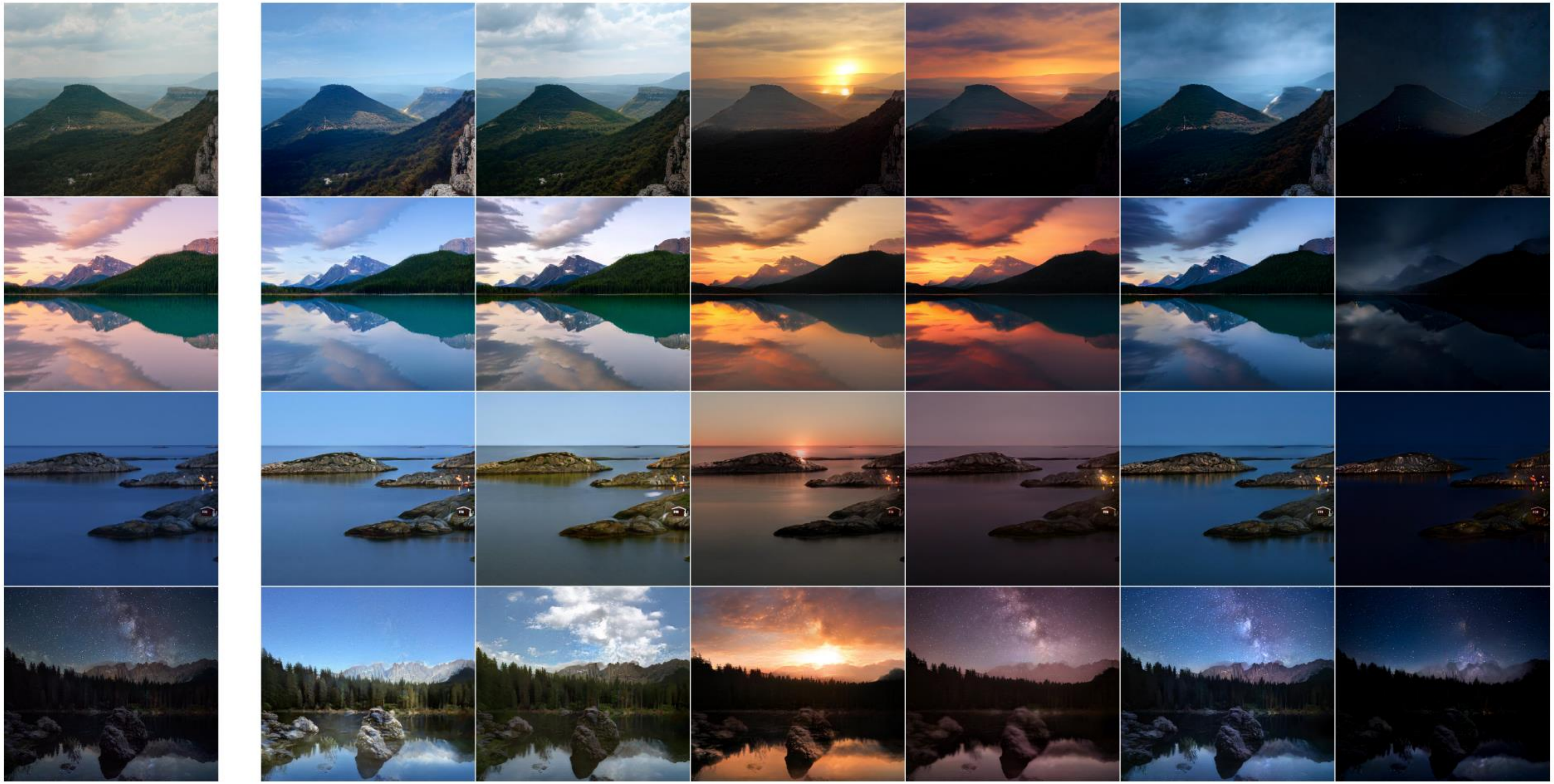
CVPR 2020 Oral

박성현

DAVIAN

Data and Visual Analytics Lab

High-Resolution Daytime Translation



Link : https://saic-mdal.github.io/HiDT/

Contribution

- Present the High-resolution Daytime Translation (HiDT) model.

- Contribution
  1) Show how to train a multi-domain image-to-image translation model on a large dataset of unaligned images without domain labels.
  2) To ensure fine detail preservation, the authors propose an architecture for image-to-image translation that combines the two well-known ideas: skip connection and adaptive instance normalizations (AdaIN)
  3) Address the task of image-to-image translation at high resolution.
     Training a high-capacity image-to-image translation network directly at high resolution is computationally infeasible. Therefore, the authors propose a new enhancement scheme that allows to apply the image-to-image translation network trained at medium resolution for high-resolution images.

Translation Network

- Content encoder E_c maps the initial image to a 3D tensor c using several convolutional downsampling layers and residual blocks.
- Style encoder E_s is a fully convolutional network that ends with global pooling and a compressing 1x1 convolutional layer.
- Generator G processes with c with several residual blocks with AdaIN modules inside and then upsamples it.
- Introduce an additional convolutional block with AdaIN and apply it to the skip connections.

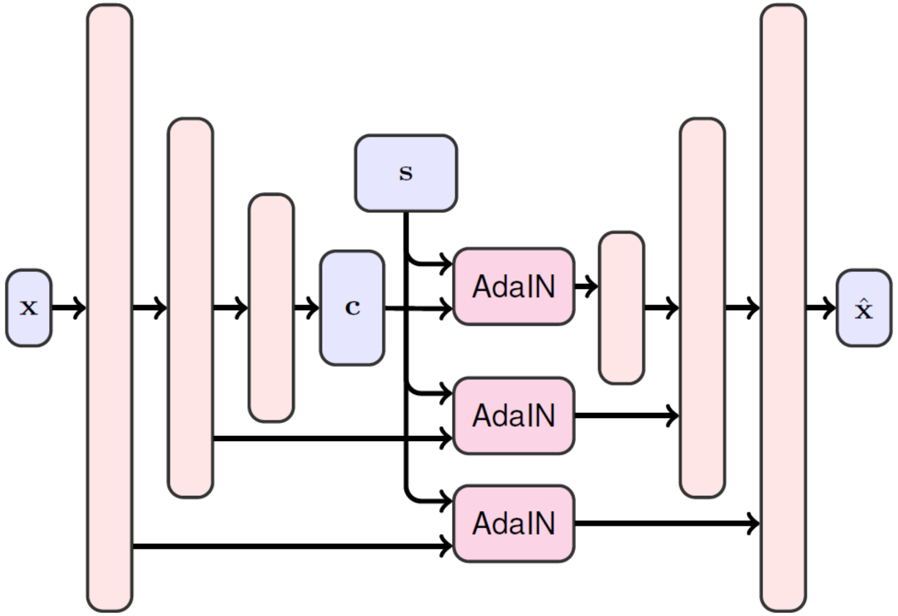
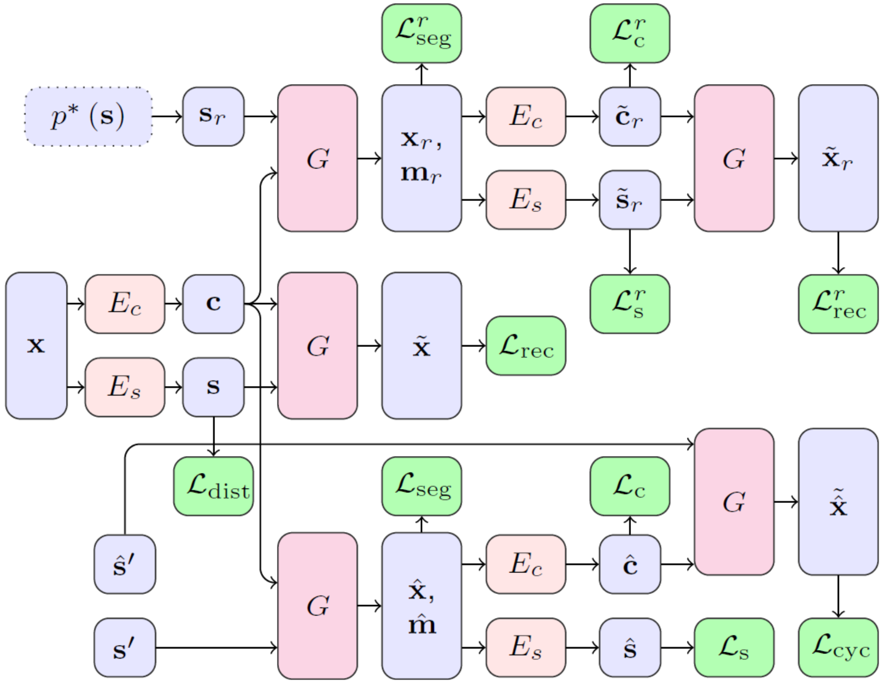


Figure 2: Diagram of the Adaptive U-Net architecture: an encoder-decoder network with dense skip-connections and content-style decomposition $(\mathbf{c}, \mathbf{s})$.

Training Loss

- Image Reconstruction Loss
  - $L_{rec} = \|\tilde{x} - x\|_1$
  - $L_{rec}^r = \|\tilde{x}_r - x_r\|_1$
  - $L_{cyc} = \left\|\tilde{\hat{x}} - x\right\|_1$
- Segmentation Loss
  - $L_{seg} = CE(m, \widehat{m})$
  - $L_{seg}^r = CE(m, m_r)$
- Adversarial Loss
  - Unconditional discriminator
  - Conditional discriminator (style vector is used as conditioning)
  - Least Square GAN



$p^*(\mathbf{s})$
$\mathbf{s}_r$
G
$\mathbf{x}_r, \mathbf{m}_r$
$\mathcal{L}^r_{\text{seg}}$
E_c
E_s
$\tilde{\mathbf{c}}_r$
$\tilde{\mathbf{s}}_r$
$\mathcal{L}^r_{\text{c}}$
$\mathcal{L}^r_{\text{s}}$
$\tilde{\mathbf{x}}_r$
$\mathcal{L}^r_{\text{rec}}$
$\mathbf{x}$
$\mathbf{c}$
$\mathbf{s}$
$\tilde{\mathbf{x}}$
$\mathcal{L}_{\text{rec}}$
$\mathcal{L}_{\text{dist}}$
$\hat{\mathbf{s}}'$
$\mathbf{s}'$
$\hat{\mathbf{x}}, \hat{\mathbf{m}}$
$\mathcal{L}_{\text{seg}}$
$\hat{\mathbf{c}}$
$\hat{\mathbf{s}}$
$\mathcal{L}_{\text{c}}$
$\mathcal{L}_{\text{s}}$
$\tilde{\hat{\mathbf{x}}}$
$\mathcal{L}_{\text{cyc}}$

Figure 3: HiDT learning data flow. We show half of the (symmetric) architecture; $\mathbf{s}' = E_s(\mathbf{x}')$ is the style extracted from the other image $\mathbf{x}'$, and $\hat{\mathbf{s}}'$ is obtained similarly to $\hat{\mathbf{s}}$ with $\mathbf{x}$ and $\mathbf{x}'$ swapped. Light blue nodes denote data elements; light green, loss functions; others, functions (subnetworks). Functions with identical labels have shared weights. Adversarial losses are omitted for clarity.

Training Loss

- Latent Reconstruction Loss
  - Cycle consistency loss with respect to the style and content codes
- Style Distribution Loss
  - To enforce the structure of the space of extracted style codes, the style distribution loss is applied to a pool of styles.
  - Style pool : $\{s^{(1)}, \dots, s^{(T)}\} \rightarrow$ mean $\hat{\mu}_s$, covariance $\hat{\Sigma}_s$
  - Style distribution loss matches empirical moments of the resulting distribution to the moments of the prior distribution $N(0, I)$
  - $L_{dist} = \|\hat{\mu}_T\|_1 + \left\|\hat{\Sigma}_T - I\right\|_1 + \left\|diag\left(\hat{\Sigma}_T\right) - 1\right\|_1$
- Total Loss
  - $\min_{Ec,E_s,G} \max_D L(E_c, E_s, G, D) = \lambda_1(L_{adv} + L^r_{adv}) + \lambda_2\left(L_{rec} + L^r_{rec} + L_{cyc}\right) + \lambda_3\left(L_{seg} + L^r_{seg}\right) + \lambda_4(L_c + L^r_c) + \lambda_5 L_s + \lambda_6 L^r_s + \lambda_7 L_{dist}$

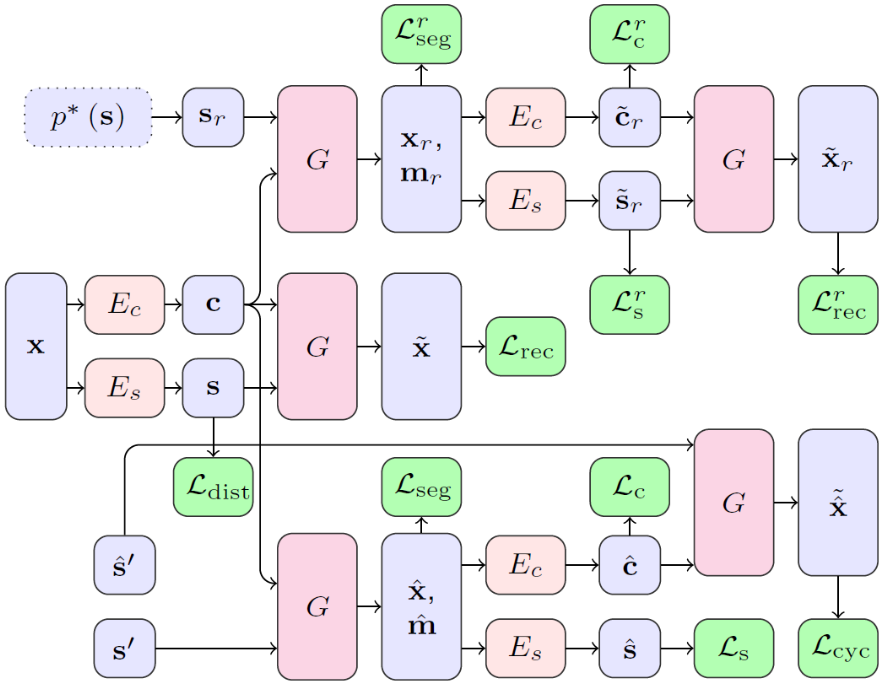
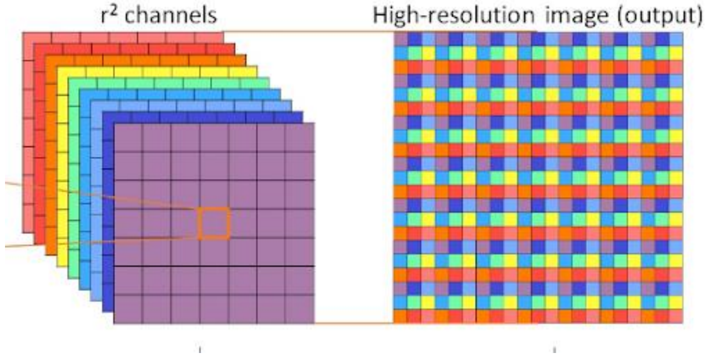


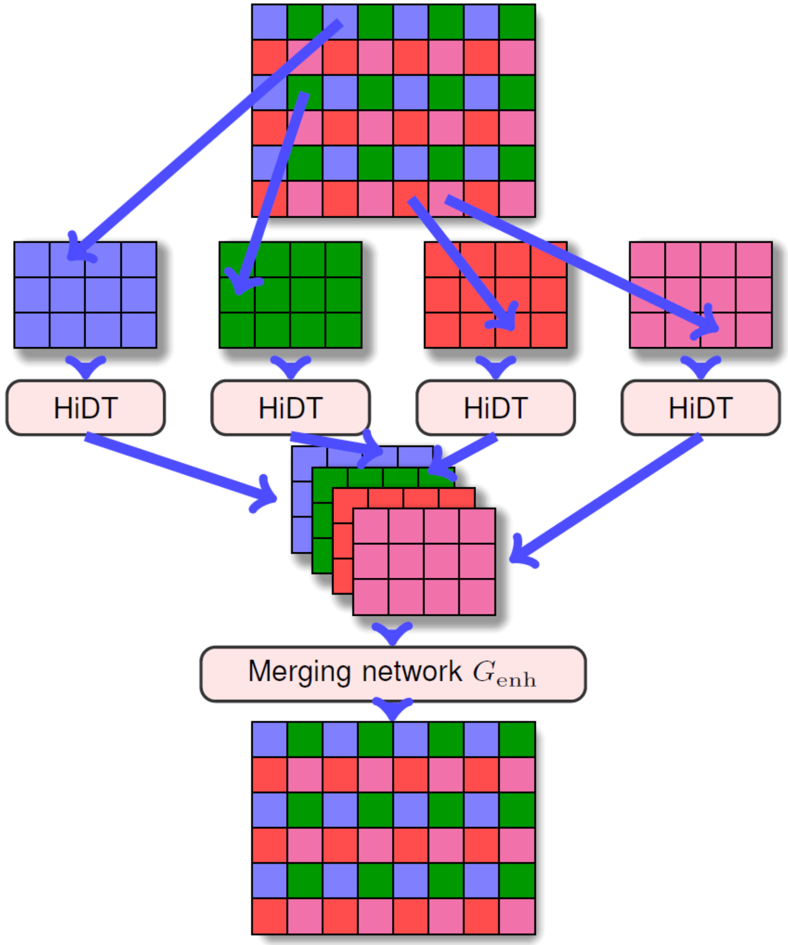
Figure 3: HiDT learning data flow. We show half of the (symmetric) architecture; $\mathbf{s}' = E_s(\mathbf{x}')$ is the style extracted from the other image $\mathbf{x}'$, and $\hat{\mathbf{s}}'$ is obtained similarly to $\hat{\mathbf{s}}$ with $\mathbf{x}$ and $\mathbf{x}'$ swapped. Light blue nodes denote data elements; light green, loss functions; others, functions (subnetworks). Functions with identical labels have shared weights. Adversarial losses are omitted for clarity.

Enhancement Postprocessing

- Apply translation multiple times at medium resolution and then use a separate merging network G_{enh} to combine the results into a high-resolution translated image.
- The merging network G_{enh} is trained in a semi-supervised mode on two datasets: paired and unpaired.
  1) To obtain each training pair, the authors take a high-res image, decompose it into 16 medium-resolution images, and pass them through the HiDT architecture without changing the style.
  2) For the unpaired dataset collection the authors use the same procedure, but the style is being sampled from normal distribution.
- During training, the authors use the same losses as *"pix2pixHD"* (perceptual, feature matching, and adversarial loss functions)



[Pixelshuffle]



[Enhancement Postprocessing]

Experiments:

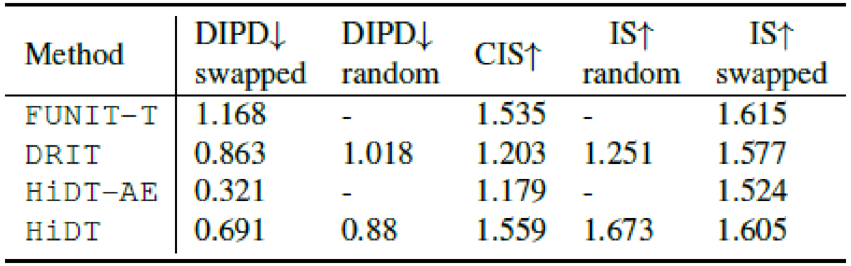
- Training details
  - The style encoder contains 4 downsampling blocks. The output of the style encoder is a 3-channel tensor, which is averages-pooled into a 3 dimensional vector.
  - The decoder has 5 residual blocks with AdaIN layers and 2 upsampling blocks.
  - AdaIN parameters are computed from the style vector via 3-layer fully-connected network.
  - For training, the images were downscaled to the resolution of 256x256.
- Dataset
  - Collected a dataset of 20,000 landscape photos from the Internet.
  - A small part of these images were manually labeled into 4 classes (night, sunset/sunrise, morning/evening, noon)
  - The authors used predicted labels (with ResNet-based classifier) in two ways:
    1) To balance the training set for image translation models with respect to daytime classes
    2) To provide domain labels for baseline models.
  - Segmentation masks were produced by an external state of the art model.

Experiments: quantitative results

- Domain-invariant perceptual distance (DIPD)
  - The L_2 distance between normalized $Conv5$ features of the original image and its translated version.
- Conditional Inception score (CIS) & Inception score (IS)
  - Measures the diversity of translation results

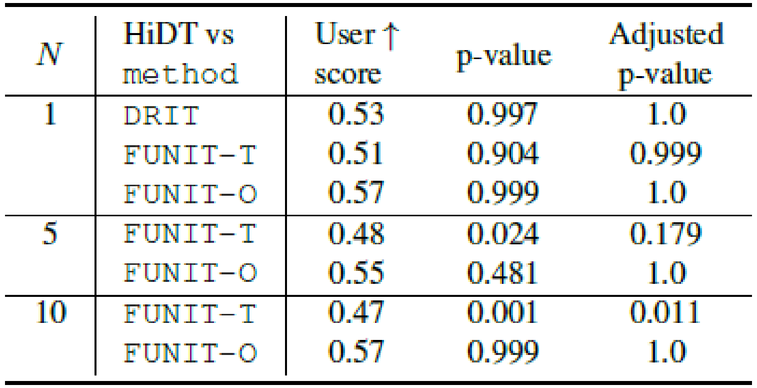
$$\text{IS} = \mathbb{E}_{x_1 \sim p(x_1)}[\mathbb{E}_{x_{1\to2} \sim p(x_{2\to1}|x_1)}[\text{KL}(p(y_2|x_{1\to2})||p(y_2))]]$$

$$\text{CIS} = \mathbb{E}_{x_1 \sim p(x_1)}[\mathbb{E}_{x_{1\to2} \sim p(x_{2\to1}|x_1)}[\text{KL}(p(y_2|x_{1\to2})||p(y_2|x_1))]]$$



Method	DIPD↓ swapped	DIPD↓ random	CIS↑	IS↑ random	IS↑ swapped
FUNIT-T	1.168	-	1.535	-	1.615
DRIT	0.863	1.018	1.203	1.251	1.577
HiDT-AE	0.321	-	1.179	-	1.524
HiDT	0.691	0.88	1.559	1.673	1.605

Table 2: Performance comparison of three models using a hold-out dataset. FUNIT is not applicable in the random setting. According to the selected metrics, none of the models shows complete superiority over the others.



N	HiDT vs method	User ↑ score	p-value	Adjusted p-value
1	DRIT	0.53	0.997	1.0
	FUNIT-T	0.51	0.904	0.999
	FUNIT-O	0.57	0.999	1.0
5	FUNIT-T	0.48	0.024	0.179
	FUNIT-O	0.55	0.481	1.0
10	FUNIT-T	0.47	0.001	0.011
	FUNIT-O	0.57	0.999	1.0

Table 1: User preference study of HiDT against the baselines. N is the number of styles averaged in the few-shot setting. The user score is the share of users that choose HiDT in the pairwise comparison. Our results show that all methods are competitive. The increase of N leads to the better quality of FUNIT-T.

Experiments: qualitative results

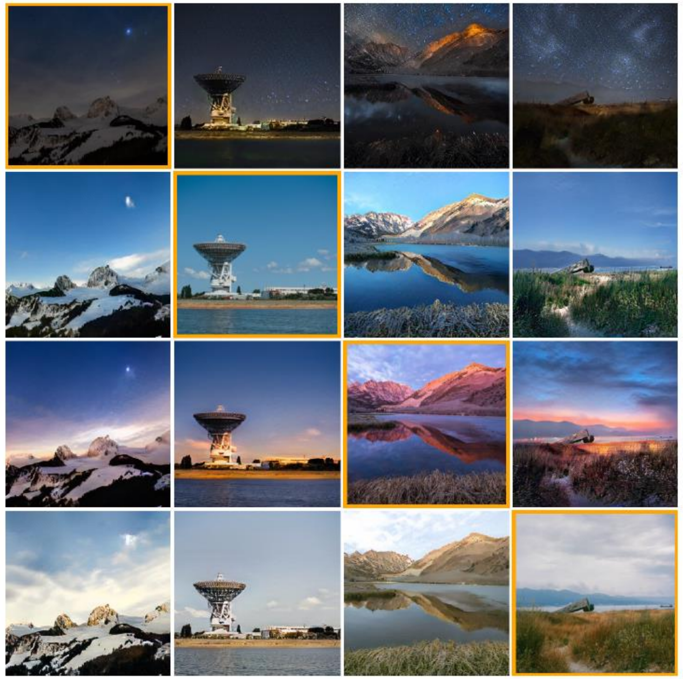


Figure 6: Swapping styles between two images. Original images are shown on the main diagonal. The examples show that HiDT is capable to swap the styles between two real images while preserving details.

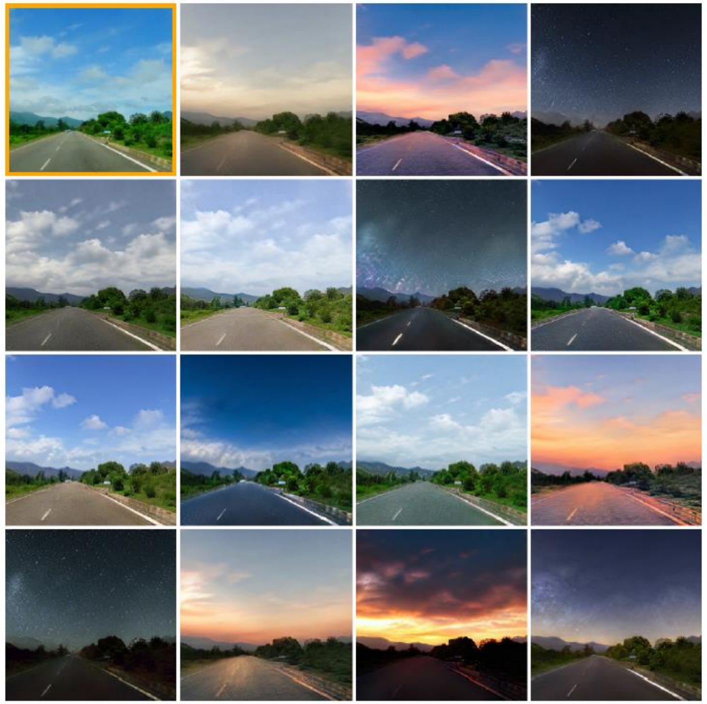


Figure 7: The original content image (top left), transferred to randomly sampled styles from prior distribution. The results demonstrate the diversity of possible outputs.

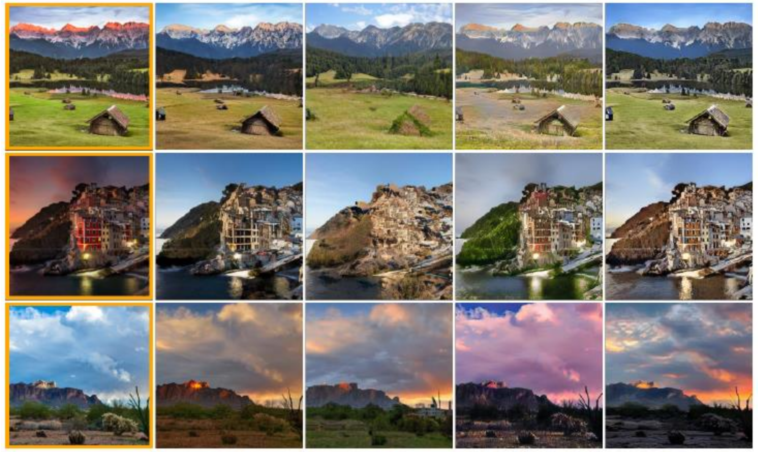


Figure 8: Comparison with baselines. Columns, left to right: the original image, FUNIT-T, FUNIT-O, DRIT, HiDT (ours). Our model, trained and applied without knowledge about domain labels, has translation quality similar to the models that require such supervision.

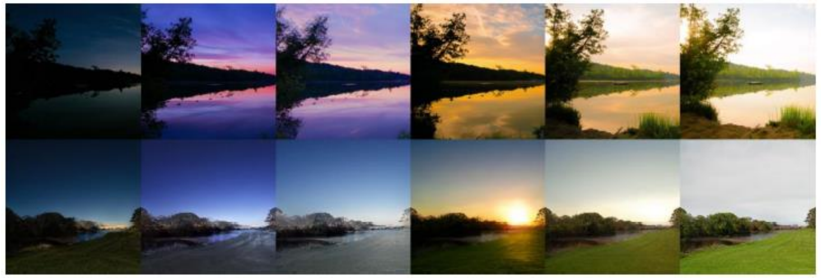


Figure 9: Timelapse generation using styles extracted from a real video. Top: frames from a guidance video. Bottom: timelapse generated from a single image using extracted styles.

Experiments: qualitative results

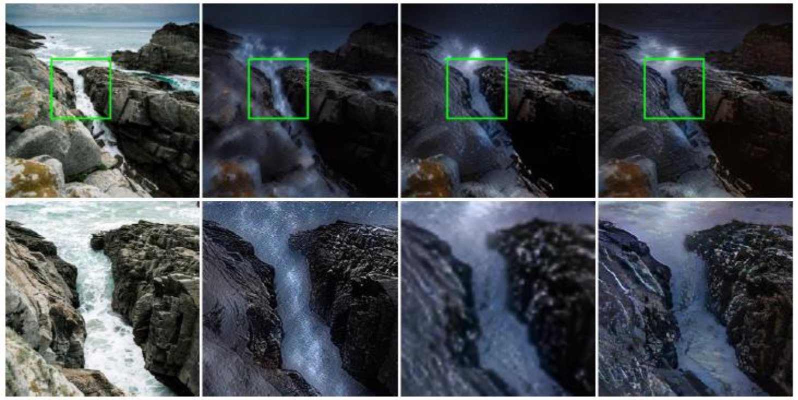


Figure 10: Enhancement of our translation network outputs with different methods. Columns, left to right: original image; result of our translation network applied directly to the hi-res input; low-res translation output upsampled with Lanczos' method; the result of our enhancement scheme. In this example, direct fully-convolutional application to hi-res turns water into sky with stars, while the enhancement network preserves the semantics of the scene.

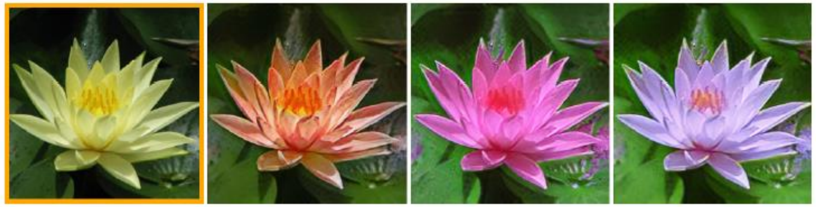


Figure 11: A flower image (left) translated to several randomly sampled styles by HiDT trained on Oxford Flowers dataset.

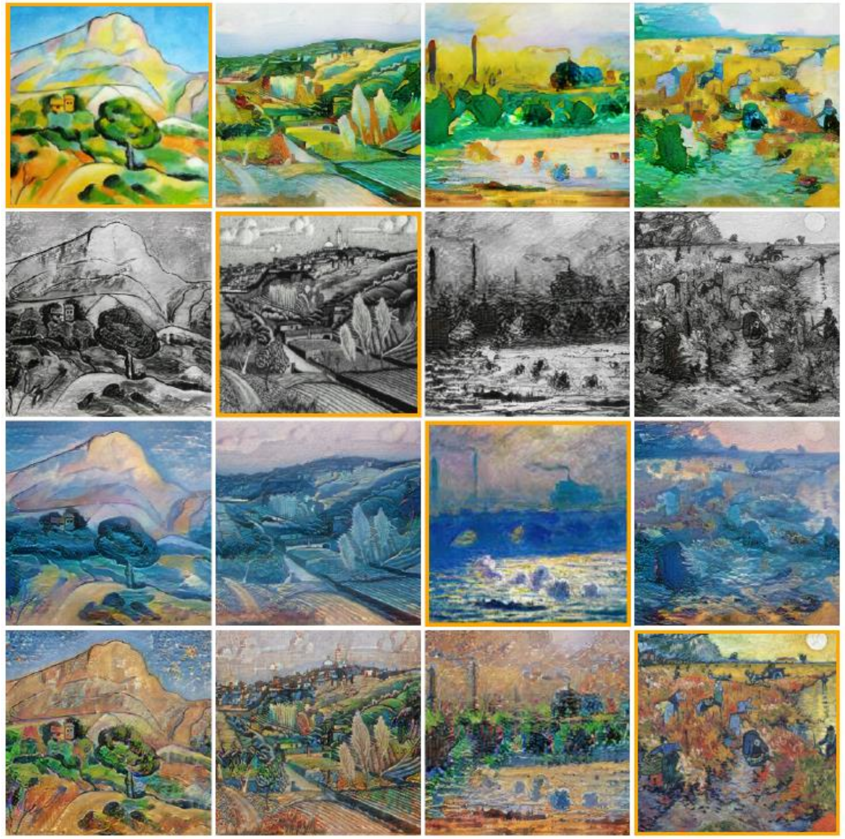


Figure 12: Style swapping for the HiDT system trained on a paintings dataset. The main diagonal contains original paintings and off-diagonal entries correspond to translated results. Plausbile translations obtained by HiDT in this case, suggests its generality.

Thank you!