

# Harnessing the Conditioning Sensorium for Improved Image Translation

Cooper Nederhood<sup>1,3</sup>, Nicholas Kolkin<sup>2</sup>, Deqing Fu<sup>1,3</sup>, and Jason Salavon<sup>1,3</sup>

<sup>1</sup>The University of Chicago   <sup>2</sup>Toyota Technological Institute at Chicago   <sup>3</sup>Jason Salavon Studio  
{cnederhood, deqing, jsalavon}@uchicago.edu, {nick.kolkin}@ttic.edu



Figure 1: Image translation results of our framework across a variety of domains and conditioning modalities

## Abstract

*Multi-modal domain translation typically refers to synthesizing a novel image that inherits certain localized attributes from a ‘content’ image (e.g. layout, semantics, or geometry), and inherits everything else (e.g. texture, lighting, sometimes even semantics) from a ‘style’ image. The dominant approach to this task is attempting to learn disentangled ‘content’ and ‘style’ representations from scratch. However, this is not only challenging, but ill-posed, as what users wish to preserve during translation varies depending on their goals. Motivated by this inherent ambiguity, we define ‘content’ based on conditioning information extracted by off-the-shelf pre-trained models. We then train our style extractor and image decoder with an easy to optimize set of reconstruction objectives. The wide variety of high-quality pre-trained models available and simple training procedure makes our approach straightforward to apply across numerous domains and definitions of ‘content’. Additionally it offers intuitive control over which aspects of ‘content’ are preserved across domains. We evaluate our method on traditional, well-aligned, datasets such as CelebA-HQ, and propose two novel datasets for evaluation on more complex scenes: ClassicTV and FFHQ-Wild. Our approach, Sensorium, enables higher quality domain translation for more complex scenes.*

## 1. Introduction

Our focus in this work is multi-modal, reference-guided, image-to-image domain translation. One well-known variant of this task is seasonal change, for example translating a photograph taken in the summer to a “winter” style, adding snow to the ground and removing leaves from trees. However, it’s not always clear which features should be invariant across domains. For example, when translating from photographic portraits to paintings a user might wish to either keep their pose or to alter it to match the canon of Renaissance painting. An additional ambiguity arises from the lack of a bijection between most domains. Even within a domain there is typically heterogeneity for a given definition of content. These challenges motivate the development of flexible domain translation models which offer control over what should be preserved as “content” **and** which mode of the target domain should be used as “style”.

Existing domain translation frameworks produce high quality results on relatively simple meta-domains like human faces, but performance deteriorates both when the complexity of a single domain increases, and when the semantic and morphological gap between translation domains grows. Rather than pose domain translation as *learning from scratch to disentangle and recombine content and style*, we instead pose it as **reconstruction from conditioning information, aided by a tightly bottle-necked ‘style’ representation**. Empirically this leads to style represen-

tations that capture only what the content does not. Content is defined by selecting conditioning modalities the user wishes to remain invariant during translation (of which a huge variety such as depth-estimation, pose-estimation, and semantic-segmentation, have been well-studied and can be inferred automatically using pre-trained models). This formulation is intuitive, and allows practitioners to select a content definition relevant to their goals.

Our framework, based solely on the well-posed task of image reconstruction, is easy to train and provides the end user with explicit and intuitive control over “content” while improving synthesis quality for complex scenes. To demonstrate the greater robustness of our pipeline on more complex domain translation tasks we introduce two novel datasets: ClassicTV and FFHQ-Wild. In addition to these complex human-centric translation datasets, we show translations from landscape photographs to different seasons and painting styles.

## 2. Related Work

Recent progress in image translation has developed as a specialized subfield of image synthesis. To date the most successful family of neural-network based image synthesis models are Generative Adversarial Networks (GANs). The discriminative loss introduced by GANs has pushed the frontier of high-quality image generation forward, and become a pillar of many image synthesis networks. Autoencoders [36], often combined with discriminative loss, have also proved vital for conditional image synthesis due to their ability to learn to extract data-driven representations from images in an unsupervised manner [51, 15, 17, 31].

While image synthesis from scratch is a challenging and scientifically interesting problem, a clear practical application of image synthesis is as a creative tool. **Conditional** image synthesis is vital in this context. Numerous forms of spatial conditioning have been explored, for example: class label [3, 25, 26, 27, 28], dense semantic segmentation [17, 18, 51, 52, 40], bounding boxes [47], and pose [11, 7, 40, 32]. Not only do conditioned models offer valuable artistic control, they also make complex image synthesis problems tractable. While remarkable progress has been made in unconditional image synthesis, the quality of results in most domains lags far behind that on aligned human faces [20]. Even high quality models trained using vast amounts of data and computational power [3, 34] are limited to datasets primarily consisting of single, large, central objects such as ImageNet [10], and suffer from a sharp drop in quality relative to models trained only on faces. However, by using ground truth segmentations from the COCO-Stuff [4] and ADE20k datasets [48], Park et al. [31] have recently proposed a method, SPADE, capable of synthesizing outputs of similar quality to [3, 34], but with significantly greater semantic and spatial complexity. Recent work by

Zhang et al. [46] has improved over SPADE, boosting synthesis quality based on ADE20k segmentations even further.

Domain translation has thus far been a specialized form of conditional image synthesis. Prior work has approached this challenge by explicitly learning to disentangle domain invariant and domain specific representations, then mixing them to synthesize a novel image [24, 30, 16], learning a bijection between domains by leveraging a cycle consistency loss [51], or by combining both approaches [52, 15]. We propose side-stepping the difficulty of learning domain invariant representations from scratch by taking advantage of automatically generated conditioning information.

Other works have explored conditional image synthesis based on dense [17, 18, 51, 52, 40, 1] and sparse [47, 7, 40, 1] spatial inputs, along with simple and domain-specific forms of automatic conditioning such as edges and face parts [46, 40]. We build on this line of work and demonstrate that even as we move beyond faces to more complex and diverse datasets, such as ClassicTV and FFHQ-Wild, pre-trained neural networks are capable of automatically generate high quality conditioning across many modalities. We leverage state-of-the-art models for depth-estimation [21], pose-estimation [2], and semantic-segmentation [12] to facilitate controllable and high-quality image synthesis.

Not learning a content representation from scratch greatly simplifies the learning process. By using a reconstruction objective we frame style extraction as simply learning to embed all information not contained in the conditioning. To force this information to be spatially invariant we take inspiration from [20, 30] and encode ‘style’ as a vector without spatial dimension extracted from a “style image”. This vector is injected into our generator as conditional de-normalization [20]. In fact, we train our domain translation framework solely based on the well-posed task of image reconstruction. This is in contrast to the more complicated training procedures of recent work, which rely on domain-translation specific regularization terms and losses applied to attempted domain translations (which typically lack a unique solution). Thus far, domain translation models have primarily focused on translation tasks driven by textural changes such as summer-to-winter or day-to-night [30, 16, 15, 8, 51, 52, 24]; or translation within spatially and semantically narrow meta-domains such as aligned human faces/portraits [43, 9, 8, 35, 24], or isolated animals on natural backgrounds [37, 30, 9, 15, 8, 24]. We propose two new human-focused datasets for evaluating image translation models: FFHQ-Wild, a dataset of human-centered crops derived from the original in-the-wild FFHQ images; along with ClassicTV, a new dataset composed of central crops from randomly sampled frames of the public domain television shows “Bonanza” and “The Lucy Show”, filtered for frames where human figures are visible. Of the image synthesis models operating in a most similar regime

to our own [11, 22, 15, 40, 1, 9, 8, 46] we compare against the most recent [9, 8, 46] and demonstrate that these models cannot perform well on more spatially complex datasets such as FFHQ-Wild or ClassicTV. We also show that unmodified SPADE [31], which leverages conditioning but is not explicitly designed with domain translation in mind, cannot perform well on these datasets; however, in our ablations we demonstrate that components of our framework can be added to SPADE to improve its performance.

### 3. Model

Our goal is to obtain a model capable of: (1) reconstruction, (2) translation within the same domain (guided by an exemplar in the same domain as the input), and (3) translation across domains (guided by an exemplar in the target domain).

Let  $X_a$  and  $Y_b$  be two domains, and  $x_a$  and  $y_b$  be images belonging to  $X_a$  and  $Y_b$  respectively. Let  $c_a = \Phi(x_a)$  be the conditioning information extracted by one (or more) pre-trained model(s) from  $x_a$ . We consider the exact form of  $\Phi$  to be a hyperparameter chosen based on the user’s goals.

There are three components of our model which we train (in contrast to the pre-trained  $\Phi$ ): a content network, a style network, and a generator network. The content network is domain-invariant, and learns to map the conditioning to a hidden representation  $\bar{c}_a = \mathcal{H}(c_a)$  useful for RGB synthesis. All spatial information is contained in  $\bar{c}_a$ .

The style network is domain-specific, and maps an RGB image to a global style vector  $\bar{s}_a = \mathcal{F}_a(y_a)$  or  $\bar{s}_b = \mathcal{F}_b(y_b)$  (depending on the domain of the style). The final output  $\hat{x}_{ab} = \mathcal{G}(\bar{c}_a, \bar{s}_b)$  is produced by the decoder network  $\mathcal{G}$  which synthesizes the final image from the spatial conditioning,  $\bar{c}_a$ , and the global style,  $\bar{s}_a/\bar{s}_b$ .

If our goal is reconstruction of  $x_a$  (an important part of training), it is produced as  $\hat{x}_{aa} = \mathcal{G}(\mathcal{H}(\Phi(x_a)), \mathcal{F}_a(x_a))$ . If our goal is style transfer within a domain, it is produced as  $\hat{x}_{aa'} = \mathcal{G}(\mathcal{H}(\Phi(x_a)), \mathcal{F}_a(x'_a))$ , where  $x'_a$  is a second image from domain  $X_a$ . If our goal is domain translation, it is produced as  $\hat{x}_{ab} = \mathcal{G}(\mathcal{H}(\Phi(x_a)), \mathcal{F}_b(y_b))$ . Note that this third regime uses a different style encoder  $\mathcal{F}_b$ , while the first two use the same style encoder  $\mathcal{F}_a$ .

#### 3.1. Training objective

While many domain translation pipelines require complex training procedures, where different sets of losses are applied to different permutations of inputs, our training procedure is very simple. We train our model entirely using content/style representations extracted from the same image, and minimize reconstruction losses in feature space along with an adversarial loss. Because reconstruction is a well-posed and easily optimized task, our model does not suffer from mode collapse, and the adversarial loss encourages the model to hallucinate plausible details not captured

by the reconstruction losses, rather than synthesize complex images from scratch.

**Feature Reconstruction Losses:** We compute our reconstruction loss in the feature spaces of two neural networks. One is pretrained VGG19, and the other is the discriminator being learned (proposed as the GAN-Stability loss [41]). We use each network to produce activations for the synthesized image and for the ground-truth image. Then, we minimize an L1 loss between hidden layers, weighting the deeper and more semantically meaningful layers more highly.

$$\mathcal{L}_{ra} = \mathbb{E}_{x_a} \sum_{i=0}^{L-1} \left[ w_i \left\| P^{(i)} \left( \mathcal{G}(\mathcal{H}(c_a), \mathcal{F}_a(x_a)) \right) - P^{(i)}(x_a) \right\| \right] \quad (1)$$

Where  $c_a = \Phi(x_a)$  is the conditioning extracted from  $x_a$ , and  $P$  is either pretrained VGG19 or  $D_a$  (the discriminator for domain  $X_a$ ).  $P^{(i)}(\cdot)$  indexes the hidden representation of  $P(\cdot)$  at layer  $i$ . Finally,  $w_i$  is the weight applied to layer  $i$ , allowing deeper layers to be weighted more highly. We define the representation loss in domain  $b$ ,  $\mathcal{L}_{rb}$ , analogously.

**Adversarial Loss:** We also use the adversarial loss proposed in Hinge-GAN [23] to encourage synthesizing domain specific details not captured by the reconstruction losses.

$$\begin{aligned} \mathcal{L}_D^a &= \mathbb{E}_{x_a} [\min(0, -1 - D_a(\mathcal{G}(\mathcal{H}(c_a), \mathcal{F}_a(x_a)))) \\ &\quad + \min(0, -1 + D_a(x_a))] \\ \mathcal{L}_G^a &= -\mathbb{E}_{x_a} [(D_a(\mathcal{G}(\mathcal{H}(c_a), \mathcal{F}_a(x_a)))] \end{aligned} \quad (2)$$

Where  $c_a = \Phi(x_a)$  is the conditioning extracted from  $x_a$ , and  $D_a$  is the discriminator for domain  $X_a$ . We emphasize that we only assess loss on the within-domain reconstruction tasks. As in the representation losses,  $\mathcal{L}_D^b, \mathcal{L}_G^b$ , are defined analogously.

#### 3.2. Architecture

**Content Encoder:** The input to the content encoder is a tensor of conditioning information  $\Phi(x_a) \in \mathbb{R}^{256 \times 256 \times d}$ , where the number of channels  $d$  varies based on chosen forms of conditioning. The content encoder,  $\mathcal{H}$ , is shared across all domains and is composed of a series of residual convolutions and downsampling layers. After each downsampling layer a feature map is emitted at the current resolution, resulting in a feature pyramid capturing the spatial and semantic information that can be extracted from the given content representation. For a more detailed description of the architecture, see Figure 2. We refer to the content/conditioning feature map at resolution  $r$  as  $\bar{c}_a^{(r)}$ , and define the set of feature maps to be  $\bar{c}_a = \{\bar{c}_a^{(r)} \mid r \in \{8, 16, 32\}\}$ .

**Style Encoder:** The style encoding networks are the only domain-specific part of our generator (there is also a separate discriminator for each domain). Given image

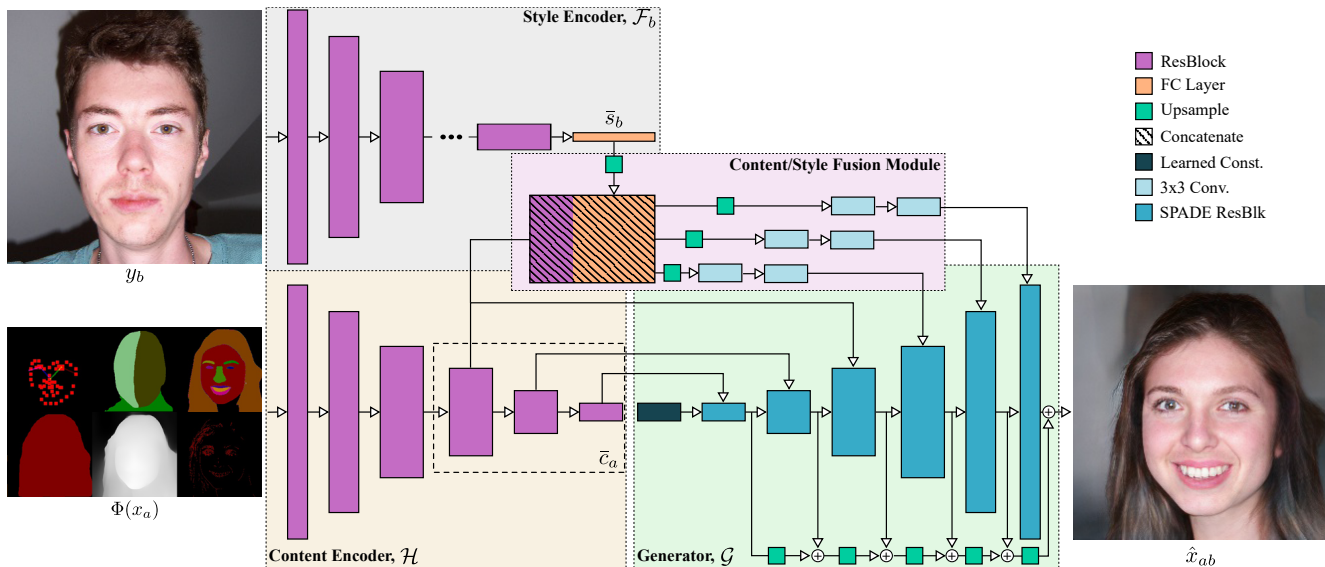


Figure 2: Overview of our **inference** framework. Note that all components of our model are trained purely for reconstruction. Translation from domain  $a$  to domain  $b$  (as seen here) is never done during training, even for more disparate domains such as FFHQ and Met-Faces.

$y_b$  in domain  $Y_b$  our style network generates latent code  $\bar{s}_b = \mathcal{F}_b(y_b)$ , and this is defined analogously for the style encoder of domain  $X_a$ . Each encoder network is composed of a series of residual blocks and pooling layers which eventually remove the spatial dimensions of the input, after which we apply a final fully-connected layer. See Figure 2 for more details.

The style encoder learns a global representation which is encouraged to contain all information about the source image not captured by the chosen conditioning sensorium (otherwise reconstruction will be impossible). While this mechanism for reconstructing content and disentangling style is largely successful, it is not perfect. We expect that our framework would benefit from future research improving reconstruction quality and integrating explicit mechanisms for disentangling style and content.

**Generator Network:** The generator network is a general purpose image synthesis module shared between domains. Agnostic to the original domain of both the source and the style reference image, the generator simply learns to fuse the content signal  $\bar{c}_a$  and the style signal  $\bar{s}_a$ . Past work [19] show that a learned global style vector can guide image synthesis by channel-wide scale and bias parameters injected via an AdaIn layer [14]. In [31] the authors introduce the SPADE injection layer, which generalizes the denormalization of AdaIn to be spatially varying function of a content embedding (itself derived from conditioning such as a semantic label map).

However, as we show in row D of Figure 3 the style of the output is a dramatically better match to the target exem-

plar if the scale/bias of SPADE injections are also a function of the style embedding. Before each SPADE injection, we fuse content and style information by spatially replicating our style embedding, then concatenating with the highest resolution feature map in feature pyramid  $\bar{c}_a/\bar{c}_b$ . This is followed by bilinear upsampling, then several convolutional layers before predicting the scale/bias parameters for SPADE injection (these layers are learned independently for each SPADE injection). We refer to this as the Content-Style Fusion Module, and it enables the generator to flexibly localize the style information.

**Discriminator Network:** We adopt the multi-resolution patch-based discriminator from SPADE [31]. The discriminator is domain-specific, and we define the discriminators for domains  $a, b$  to be  $D_a, D_b$  respectively.

## 4. Evaluation

Typical domain translation benchmarks include domain pairs like photograph to painting and summer to winter which highlight textural variations rather than complex morphological differences. We demonstrate performance on these textural-based translation tasks (see Figure 4 and Figure 7) and methodically compare on multiple morphologically complex tasks. We benchmark our method on the CelebA-HQ Male/Female datasets and also introduce two novel datasets with much greater spatial variety. We show that performance decays as the spatial complexity increases. In Figure 5 we visualize the relative spatial complexity of CelebA-HQ and our two new datasets: FFHQ-Wild Male/Female and ClassicTV Bonanza/LucyShow.



Figure 3: Ablation between our method and baseline (SPADE) [31]. (A) is SPADE as originally proposed, this synthesizes high quality images when using semantic segmentation as conditioning. (B) is identical to (A) except the conditioning is sparse facial keypoints and head pose, the baseline’s performance with sparse conditioning deteriorates significantly. (C) introduces our content encoder,  $\mathcal{H}$ , before the SPADE residual blocks and image quality approaches that of the original dense conditioning. (D) replaces the VAE based style encoder of [31] with our proposed style encoder architecture (shared between domains),  $\mathcal{F}$ , and Style-Content Fusion module, this dramatically improves how well the output matches the target style. Finally, (E) introduces a separate style encoder and discriminator for each domain ( $\mathcal{F}_a, \mathcal{F}_b$ , and  $\mathcal{D}_a, \mathcal{D}_b$ ). This emphasizes features characteristics particular to a given domain. For example, note that only (E) recreates the red lips from style reference 3. (E) is our proposed model, Sensorium.

#### 4.1. Datasets

**CelebA-HQ Male/Female** A dataset of 11,057 high resolution male faces and 18,943 female faces. We hold out 1,000 male faces and 1,000 female faces for validation.

**FFHQ-Wild Male/Female** The original FFHQ dataset encompasses 70,000 images of closely cropped human faces. To create a dataset of complex human poses we instead download the original raw in-the-wild images and

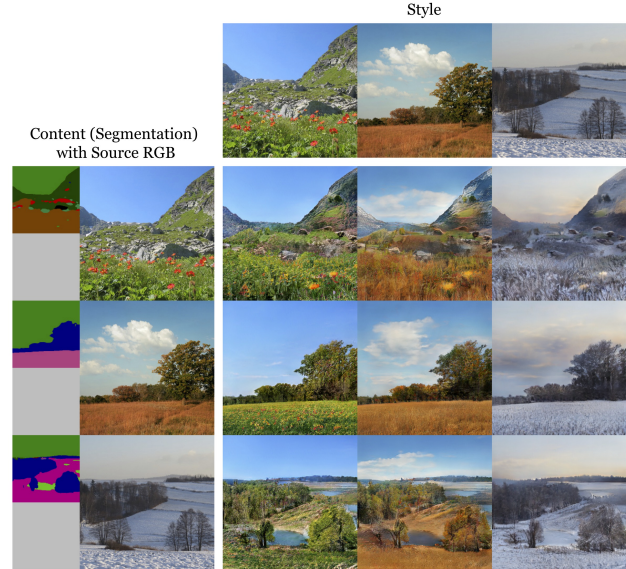


Figure 4: Weight sharing in Sensorium allows for efficient training of multidomain models, like translating between Autumn, Spring, and Winter.



Figure 5: Complexity visualization of evaluation datasets. A random sample from each dataset is shown alongside the mean of all images in that dataset. Row 1 shows CelebA-HQ male and female. Row 2 shows FFHQ-Wild male and female. Finally, row 3 shows ClassicTV Bonanza and LucyShow

then use an off the shelf Mask-RCNN model to identify 1-2 human figures from each raw image. We acquire noisy gender labels from the recent FFHQ-Aging dataset [29]. The procedure yields a dataset of 23,462 male crops and 27,305 female crops, from which hold out 238 male crops and 269 female crops for validation.

**ClassicTV Bonanza/LucyShow** We also publish a new dataset composed of human figure crops from two classic

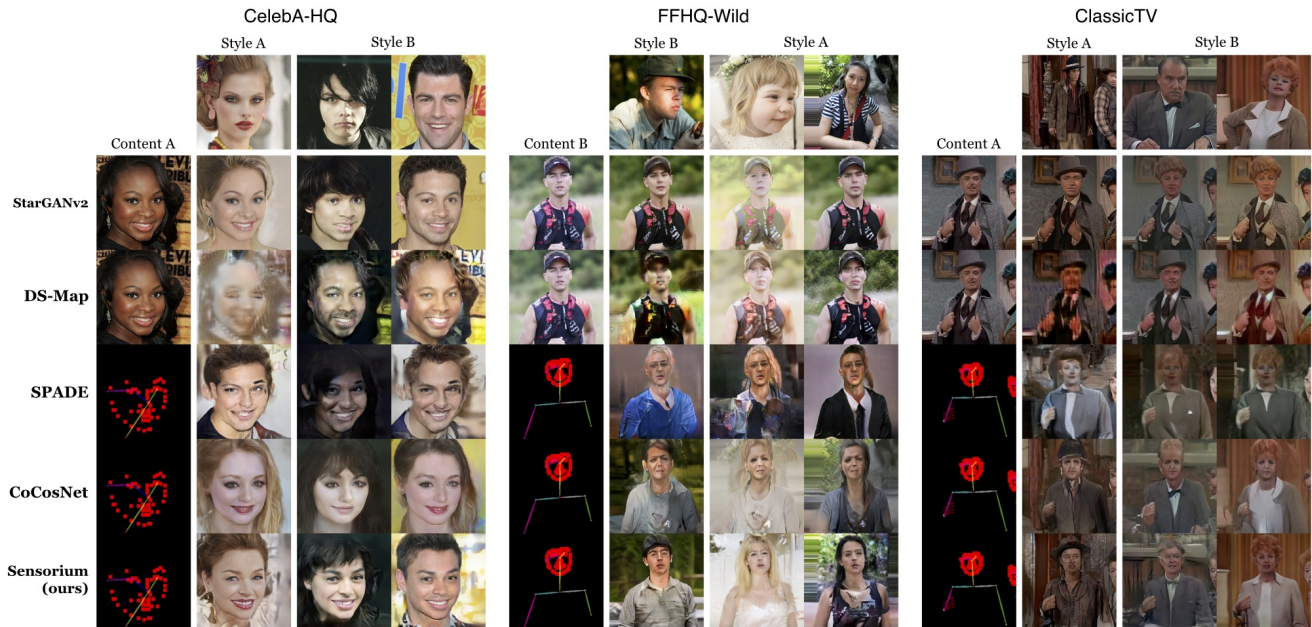


Figure 6: Qualitative comparison of our model and state-of-the-art domain translation models, focusing on the ability to make spatial deformations as dataset complexity increases. From left to right, we picture a content image and then three style reference images. Below the style reference images are synthesized images from StarGANv2 [9], DS-Map[8], SPADE[31], CoCosNet[46], and our Sensorium method. We show results for the common task of CelebA-HQ male-female translation where the domains are spatially aligned. We also show results for the spatially varied FFHQ-Wild and ClassicTV datasets. While all methods degrade in quality from CelebA-HQ to FFHQ-Wild and ClassicTV, Sensorium maintains the highest synthesis quality while matching the style well (unlike StarGANv2 and DS-Map, which perform a palette shift).

American television shows: Bonanza and The Lucy Show. Both shows have episodes which have entered the public domain and for each show we randomly sample 40,000 crops from the set of public episodes. To extract a crop from a given frame we use the same process as in FFHQ-Wild, using an off the shelf Mask-RCNN model to identify and crop human figures. For each show, we reserve 10,000 crops sampled from held out episodes.

**Flickr Seasons** We qualitatively illustrate performance on the seasonal change problem by downloading season-specific landscape images from Flickr. For Spring, Autumn, and Winter we download approximately 4,000 public images for each domain and we reserve 10% for testing purposes.

**WikiArt Landscapes** We use 16,000 landscape paintings obtained via the public WikiArt database to learn translation from photographic to painterly landscapes. We reserve 1,000 of these images for testing purposes. [39]

**Derived Conditioning Data** For each dataset we precompute automatically generated conditioning using multiple SOTA models, including: depth [33], human pose [5] [38] [6] [42], COCO-based segmentations [44] [12], ADE-based semantic segmentations [50] [49], facial segmentations [45], and human body part segmentations [2].

## 4.2. Baseline Methods

As discussed, the specifics of domain translation are inherently task dependent. Textural-based translation like Seasonal Change and Photo to Painting are often addressed by style-transfer models whereas more morphological translations like male face to female face and rendering from human pose are often addressed by exemplar-guided image synthesis models. Because we wish to show the robustness of our Sensorium approach to a variety of tasks we compare to both model paradigms, using state-of-the-art approaches for each.

First, as discussed, most domain translation models learn the disentangled style and content representations directly from the RGB images. Originally proposed in MUNIT [15], more recent work StarGANv2 [9] and DS-Map [8] show improved results. StarGANv2 allows for a style-reference image and sets state-of-the-art performance on the CelebA-HQ benchmark. DS-Map, a style transfer model, maps the content images to a shared space, then further maps the latent representation to domain-specific content spaces, allowing for more expressive content. DS-Map shows performance on a variety of tasks including on the animal faces dog/cat benchmark, which requires morphological changes.

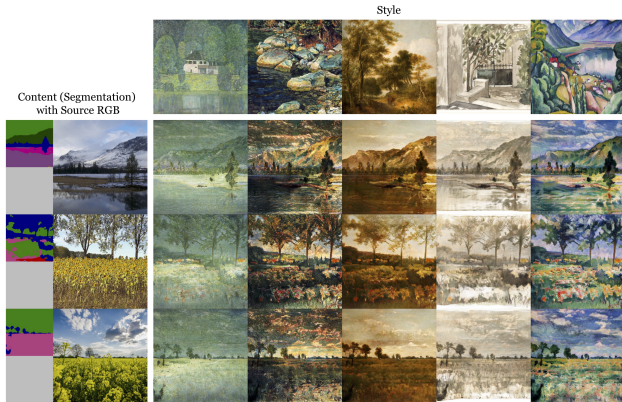


Figure 7: We illustrate performance on the Photos to Paintings task.

Under our Sensorium approach we allow the choice of derived conditioning to fix the content, allowing style to extract all other image characteristics. Therefore, we also compare Sensorium against models which perform conditional image synthesis with style control via a reference image. SPADE [31], which serves as the starting point for Sensorium, allows for style injection via a reference image. More recently, CoCosNet [46] shows impressive results on exemplar-based image synthesis using a variety of conditioning types including dense semantic segmentations, body pose, and edges.

### 4.3. Qualitative Comparison

Existing methods achieve high-fidelity domain translation when the domains are well aligned. Dataset alignment provides a strong spatial prior for translation. Without this alignment these models lose the ability to make meaningful transformations across domains, and their output devolves to simply palette-matching the style reference.

In contrast, our method can map complex textural and morphological characteristics from a style reference, while adhering to the layout and semantics provided by the conditioning information. Our approach is more robust to large pose and morphological changes. In Figure 6 we show how our approach and prior work behave as the morphological differences between the translation domains grows. Further, while DS-Map and StarGANv2 do not *explicitly* use conditioning data we show that their performance deteriorates without well-aligned data. This well-aligned data (like for human faces) can only be obtained at scale by employing a pre-trained facial recognition model, essentially the very same models we use to obtain our derived conditioning data. Therefore, our Sensorium approach essentially turns this *implicit* data-processing use of conditioning data into an *explicit* network use of conditioning data. In doing so, our approach becomes more robust to a variety of translation tasks. Please see the supplement for additional results.

### 4.4. Quantitative Comparison

We train each of our three comparison models on our three domain translation datasets: CelebA-HQ M/F; FFHQ-Wild M/F; and ClassicTV. We then train our method using human pose estimated via OpenPose as our initial content representation. We use the Fréchet inception distance (FID) to quantify performance on the domain transfer task [13]. To generate the inference population of synthesized images, given a source domain providing content and a (different) target domain providing style reference, we sample 100 images from the validation population for each. We generate all pairwise combinations of content and style reference to yield a total of 10,000 images. Following [9], we use the training population from the target domain as our ground-truth population. We report the FID figures in Table 1.

	FID ↓				
	DS-Map	StarGAN-v2	SPADE	CoCosNet	Ours
(Celeb-A) F → M	37.37	<b>32.05</b>	46.56	78.39	38.69
(Celeb-A) M → F	<b>23.81</b>	24.81	37.10	68.99	25.33
(FFHQ-Wild) F → M	54.41	84.71	121.22	74.74	<b>40.10</b>
(FFHQ-Wild) M → F	61.04	121.21	152.59	98.32	<b>45.92</b>
Bonanza → Lucy	85.25	83.39	137.18	60.70	<b>57.12</b>
Lucy → Bonanza	103.71	83.49	130.74	67.44	<b>47.62</b>
Average	60.93	71.61	104.23	74.76	<b>42.46</b>

Table 1: Quantitative comparison. We use FID scores (lower is better) to measure the quality of each method on image style translation. Red texts indicate the best method.

### 4.5. Control Via Choice of Content Representation

Our method reliably constrains the synthesized output to match the conditioning information chosen as training inputs. This allows a user to choose the content representation before training (i.e. conditioning information) that suits their goal. For example, a classic style transfer task like transferring the texture of a painting to a photograph calls for holding all geometry constant. In contrast gender transformation tasks typically call for some morphological changes, and conditioning using sparse facial landmarks may be more appropriate. See Figure 8 for detailed examples of how we achieve different degrees of content preservation through the choice of the derived conditioning.

## 5. Conclusion

While interesting from a purely scientific perspective, we predict image translation will see rapid growth in usage as a creative tool as user controls are made available. This work explores the use of pretrained models to generate conditioning data at various levels of abstraction, and effectively harnesses the outputs in an image translation system.

We show that using automatically generated conditioning allows Sensorium to synthesize compelling domain translations between spatially complex and unaligned do-



Figure 8: Illustration of the effect on Sensorium of conditioning information with varying levels of detail. Our style encoder learn to extract information needed for reconstruction not contained by the conditioning. As conditioning becomes more detailed the style’s effect on the final output decreases. For example, with facial keypoints selected as conditioning Sensorium extracts all remaining geometry including hair and glasses from the style reference image. Adding depth estimation to conditioning then restricts the amount geometry inherited from the style image - hair outline no longer reflects the reference image. Finally, adding facial segmentations fixes features like glasses and edges fix high frequency details like facial hair.

mains. In addition, our framework is able to gracefully adjust to different forms of conditioning, filling in aspects of the output the conditioning does not specify using the style of a target exemplar.

However, tension between identity preservation and domain translation remains a challenge. In particular, our framework can use conditioning data in conflict with the target domain. Details like ties, beards, and hair will be rendered in the new style whether or not it is appropriate for the target domain. An avenue for future work to address this is developing methods which translate the conditioning infor-

mation itself or which integrate global content constraints such as identity.

Other avenues for future work include extending Sensorium for video translation and a deeper exploration of the relationship between conditioning modalities and corresponding ‘null-space’ controlled by style.

**Acknowledgements** We thank Greg Shakhnarovich for helpful discussion and Jan Brugger for data compilation. This research was supported by the Center for Data and Computing at the University of Chicago and Jason Salavon Studio.



## References

- [1] Badour AlBahar and Jia-Bin Huang. Guided image-to-image translation with bi-directional feature transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9016–9025, 2019. 2, 3
- [2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2, 6
- [3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2
- [4] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 2
- [5] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 6
- [6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 6
- [7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5933–5942, 2019. 2
- [8] Hsin-Yu Chang, Zhixiang Wang, and Yung-Yu Chuang. Domain-specific mappings for generative adversarial style transfer. *arXiv preprint arXiv:2008.02198*, 2020. 2, 3, 6
- [9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 2, 3, 6, 7
- [10] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2
- [11] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 2, 3
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 6
- [13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 7
- [14] Xun Huang and Serge J Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. 2017. 4
- [15] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 2, 3, 6
- [16] 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. 2
- [17] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016. 2
- [18] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Manipulating attributes of natural scenes via hallucination. *arXiv preprint arXiv:1808.07413*, 2018. 2
- [19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 4
- [20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2
- [21] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019. 2
- [22] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 3
- [23] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 3
- [24] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 2
- [25] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018. 2
- [26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2
- [27] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018. 2
- [28] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651, 2017. 2
- [29] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *European Conference on Computer Vision*, pages 739–755. Springer, 2020. 5
- [30] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 2
- [31] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Com-*

- puter Vision and Pattern Recognition, pages 2337–2346, 2019. 2, 3, 4, 5, 6, 7
- [32] Shengju Qian, Kwan-Yee Lin, Wayne Wu, Yangxiaokang Liu, Quan Wang, Fumin Shen, Chen Qian, and Ran He. Make a face: Towards arbitrary high fidelity face manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10042, 2019. 2
- [33] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019. 6
- [34] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019. 2
- [35] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020. 2
- [36] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985. 2
- [37] Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. *arXiv preprint arXiv:2007.07431*, 2020. 2
- [38] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multi-view bootstrapping. In *CVPR*, 2017. 6
- [39] Wei Ren Tan, Chee Seng Chan, Hernán E Aguirre, and Kiyoshi Tanaka. Artgan: Artwork synthesis with conditional categorical gans. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3760–3764. IEEE, 2017. 6
- [40] Miao Wang, G. Yang, R. Li, Run-Ze Liang, Song-Hai Zhang, P. M. Hall, and S. Hu. Example-guided style-consistent image synthesis from semantic labeling. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1495–1504, 2019. 2, 3
- [41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan 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. 3
- [42] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 6
- [43] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Disentangling content and style via unsupervised geometry distillation. *arXiv preprint arXiv:1905.04538*, 2019. 2
- [44] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 6
- [45] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 6
- [46] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 2, 3, 6, 7
- [47] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019. 2
- [48] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2
- [49] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6
- [50] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal on Computer Vision*, 2018. 6
- [51] 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. 2
- [52] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017. 2