Guided Image-to-Image Translation with Bi-Directional Feature Transformation

Badour AlBahar Jia-Bin Huang
Virginia Tech
https://github.com/vt-vl-lab/Guided-pix2pix

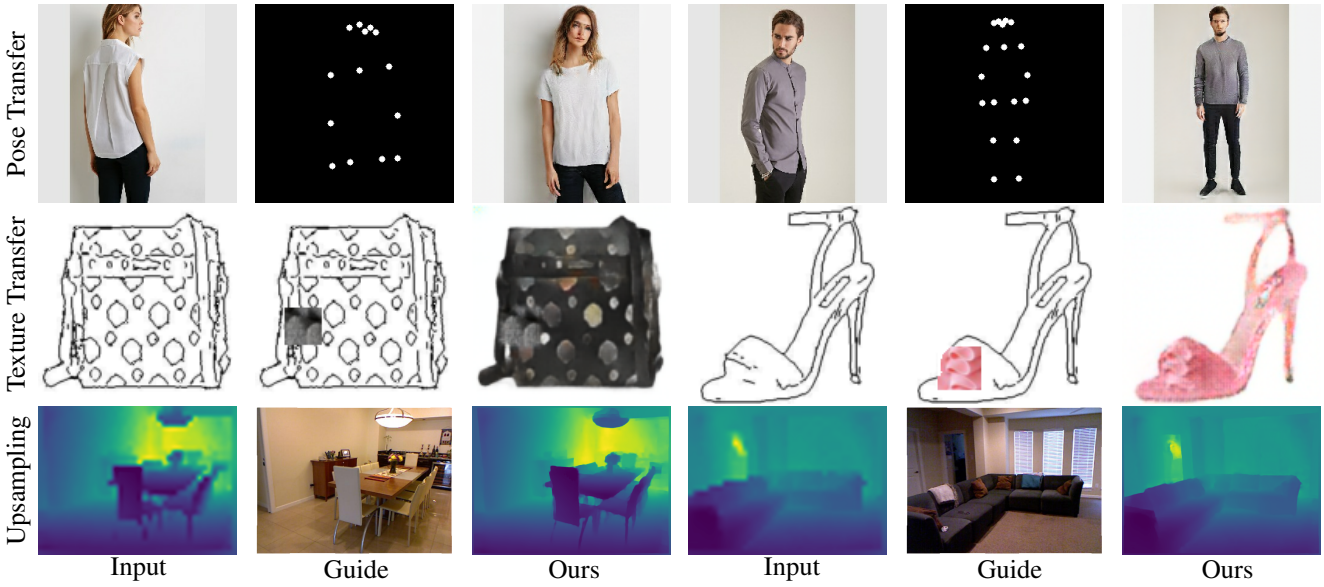


Figure 1. **Applications of guided image-to-image translation.** We present an algorithm that translates an input image into a corresponding output image while respecting the constraints specified in the provided guidance image. These controllable image-to-image translation problems often require task-specific architectures and training objective functions as the guidance can take various different forms (e.g., color strokes, sketch, texture patch, image, and mask). We introduce a new conditioning scheme for controlling image synthesis using available guidance signals and demonstrate applicability to several sample applications, including person image synthesis guided by a given pose (*top*), sketch-to-photo synthesis guided with a texture patch (*middle*), and depth upsampling guided with an RGB image (*bottom*).

Abstract

We address the problem of guided image-to-image translation where we translate an input image into another while respecting the constraints provided by an external, user-provided guidance image. Various conditioning methods for leveraging the given guidance image have been explored, including input/feature concatenation and conditional affine transformation of feature activations. All these conditioning mechanisms, however, are uni-directional, i.e., no information flow from the input image back to the guidance. To better utilize the constraints of the guidance image, we present a bi-directional feature transformation (bFT) scheme. We show that our bFT scheme outperforms other conditioning schemes and has comparable results to state-of-the-art methods on different tasks.

1. Introduction

In an image-to-image translation problem [17], we aim to translate an image from one domain to another. Many problems in computer vision, graphics, and image processing can be formulated as image-to-image translation tasks, including semantic image synthesis, style transfer, colorization, sketch to photos, to name a few. An extension to these image-to-image translation problems involves an additional *guidance image* that helps achieve controllable translation. A guidance image typically reflects the desired visual effects or constraints specified by a user or provides additional information via other modalities (color/depth, flash/non-flash, color/IR). A guidance image can thus take many different forms, e.g. color strokes or palette, semantic labels, texture patch, image, or mask. As such, most of the existing solutions for such problems often have application-

specific architectures and objective functions, and consequently cannot be directly applied to other problems.

The main technical question for guided image-to-image translation problems is how the conditional guidance image is used to affect the processing of the input source image. Various forms of conditioning schemes have been proposed in the literature. The most common one is to directly concatenate the input source image and the guidance image at the input level (i.e., concatenation along the channel dimension). While being parameter efficient, this approach assumes that the additional guidance is required at the input level and the information can be carried through all the subsequent layers. Another commonly used alternative is to concatenate the guidance and the input information at the feature level, assuming that the guidance feature representation is required at a certain level within the model.

A recent generalized conditioning scheme formalized as Feature-wise Linear Modulation (FiLM) has been successfully applied in visual reasoning task [32]. In this scheme, affine transformations are applied to intermediate feature activations using scaling and shifting parameters learned from some external conditional information. In this approach, the learned scaling and shifting operations are applied *feature-wise* (i.e., spatially invariant). There are other conditioning approaches similar to FiLM that have shown effectiveness in the context of style transfer. In this task, given an input image and a guidance style image, the goal is to synthesize an image that combines the content of the input image with the style of the guidance image. One such approach is conditional instance normalization (CIN) [7], which can be seen as a FiLM layer replacing a normalization layer. In CIN, the feature representation is first normalized to zero mean and unit standard deviation. Then an affine transformation is applied to the normalized feature representation using scaling and shifting parameters learned from the guidance style image. Another approach is adaptive instance normalization (AdaIN) [14]. AdaIN is very similar to CIN, however, unlike CIN, it does not learn the affine transformation parameters but uses the mean and standard deviation of the guidance style image as the scaling and shifting parameters respectively.

In this work, we propose a generalized conditioning scheme to incorporate the guidance image into the image-to-image translation model and show its applicability to different applications. There are two key differences between our proposed approach and the existing conditioning schemes. First, we propose to apply the conditioning operation in *both* direction with information flowing not only from the guidance image to the input image, but from the input image to the guidance image as well. Second, we extend the existing feature-wise feature transformation to be *spatially varying* to adapt to different contents in the input image. We refer to our proposed approach as bi-directional feature transformation (bFT). We validate the design of bFT through extensive experiments across multiple applications, including pose guidance appearance transfer, image synthesis with texture patch guidance, and joint depth upsampling. We demonstrate that our method, while not application-specific, achieves competitive or better performance than the state-of-the-art. Through extensive ablation study, we also show that the proposed bFT is more effective than commonly used conditional schemes such as input/feature concatenation, CIN [7] and AdaIN [14].

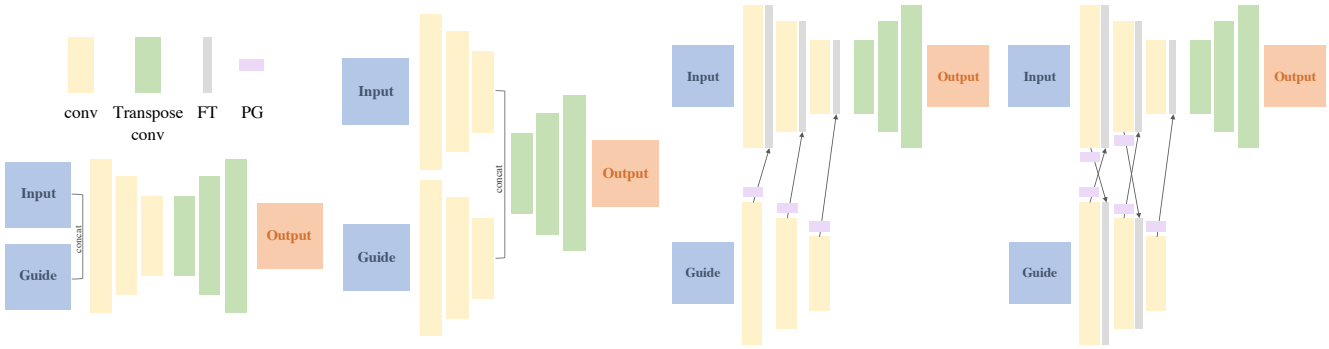
We make the following two contributions. First, we present the *bi-directional* feature transformation for generic guided image-to-image translation tasks. Compared to existing approaches that only allow the information flow from guidance to the source image, we show that incorporating the information from the input to the guidance further help improve the performance of the end task. Second, we propose a *spatially varying* extension of feature-wise transformation to better capture local contents from the guidance and the source image.

2. Related Work

Image-to-image translation A generative model is an approach to learn a data distribution to generate new samples. One widely used technique is generative adversarial networks (GANs) [9]. In GANs, there is a generator that tries to generate samples that look realistic to fool the discriminator, which tries to accurately tell whether a sample is real or fake. Conditional GANs extend the GANs by incorporating conditional information. One specific application of conditional GANs is image-to-image translation [17, 36, 31]. Several recent advances include learning from unpaired dataset [42, 38, 25], improving diversity [20, 15, 43], application to domain adaptation [2, 13, 4], and extension to video [35].

Our work builds upon the recent advances in image-to-image translation and aims to extend it to a broader set of controllable image synthesis problems. We develop our network architecture similar to that of the pix2pix [17], but the proposed bi-directional and spatially varying feature transformation layer is network-agnostic.

Guided image-to-image translation A variant of image-to-image translation problem is to incorporate additional guidance image. In a guided image-to-image translation problem, we aim to translate an image from one domain into another while respecting certain constraints specified by a guidance image. This guidance image can take many forms. Examples include color strokes [21, 27], patches [41], or color palette [3] to aid in user-guided colorization. The guidance can also be a domain label, as in a multi-domain image-to-image translation [5]. Another form could be a style image as in the problem of style transfer [7, 8, 14],



(a) Input Concatenation (b) Feature Concatenation (c) Uni-directional FT (d) Ours: Bi-directional FT

Figure 2. **Conditioning schemes.** There are many schemes to incorporate the additional guidance into the image-to-image translation model. One straight forward scheme is (a) input concatenation, this will assume that we need the guidance image at the first stage of the model. Another scheme is (b) feature concatenation. It assumes that we need the feature representation of the guide before upsampling. In (c) we replace every normalization layer with our novel feature transformation (FT) layer that manipulates the input using scaling and shifting parameters generated from the guide using a parameter generator (PG). We denote this uni-directional scheme as uFT. In this work, we propose (d) a bi-directional feature transformation scheme denoted as bFT. In bFT, the input is manipulated using scaling and shifting parameters generated from the guide and the guide is also manipulated using scaling and shifting parameters generated from the input.

a texture patch to texturize a sketch image [37], or a high-resolution RGB image to aid in depth upsampling [24, 23]. Moreover, the guidance signal could be the multi-channel and sparse, such as pose landmark for pose guided person image synthesis problems [28, 29, 33, 30]. The guidance could also be a mask and sketch enabling users to inpaint and manipulate images [39]. Due to the many different possible forms of the guidance images, most of the existing solutions for this class of problems are tailored toward specific applications, e.g., with specifically designed network architectures and training objectives.

Compared to many existing efforts in guided image-to-image translation, we focus on developing a conditioning scheme that is *application-independent*. This makes our technique more widely applicable to many tasks with different forms of guidance.

Conditioning schemes Figure 2 compares with several commonly used conditioning schemes. The most straightforward way of performing guided image-to-image translation is to concatenate the input and the guidance image (along the feature channel dimension), followed by conventional image-to-image translation models. Such an input concatenation approach can be viewed as a simple conditioning scheme. This approach assumes that the guidance signals are required from the input stage [39, 41, 37]. Several other types of conditioning schemes have been proposed in the literature. Instead of concatenating the guidance and the input image at the input, one can also concatenate their feature activations at a certain layer [23, 19]. However, it may be non-trivial to choose a suitable level of the layer to concentrate input/guidance features for subsequent processing. A recent and a more general scheme, Feature-wise Linear Modulation (FiLM) [32], applies feature-wise affine transformation using scaling and shifting parameters generated from conditioning information. Such a scheme has shown improved performance when applied to the problem of visual reasoning. Other variations of FiLM have shown good performance in the context of style transfer. Those approaches can be seen as replacing a normalization layer with a FiLM layer. One notable approach is the conditional instance normalization (CIN), where the scaling and shifting parameters are learned [7]. Another approach is adaptive instance normalization (AdaIN) where instead of learning the scaling and shifting parameters, the mean and standard deviation from the guidance features are used directly [14].

Unlike existing conditioning schemes that allow information flow only from the guidance to the input (i.e., uni-directional conditioning), we show that the proposed *bi-directional conditioning* method leads to sizable performance improvement. Furthermore, we generalize the existing spatially invariant feature-wise transform methods to support *spatially varying* transformation.

3. Bi-Directional Feature Transformation

In this work, we aim to translate an image from one domain to another while respecting the constraints specified by a given guidance image. To tackle this problem, we propose Bi-Directional Feature Transformation (bFT) to incorporate the additional guidance image into the conditional generative model. We show that this conditioning scheme can be applied to various guided image-to-image translation problems without application-specific designs.

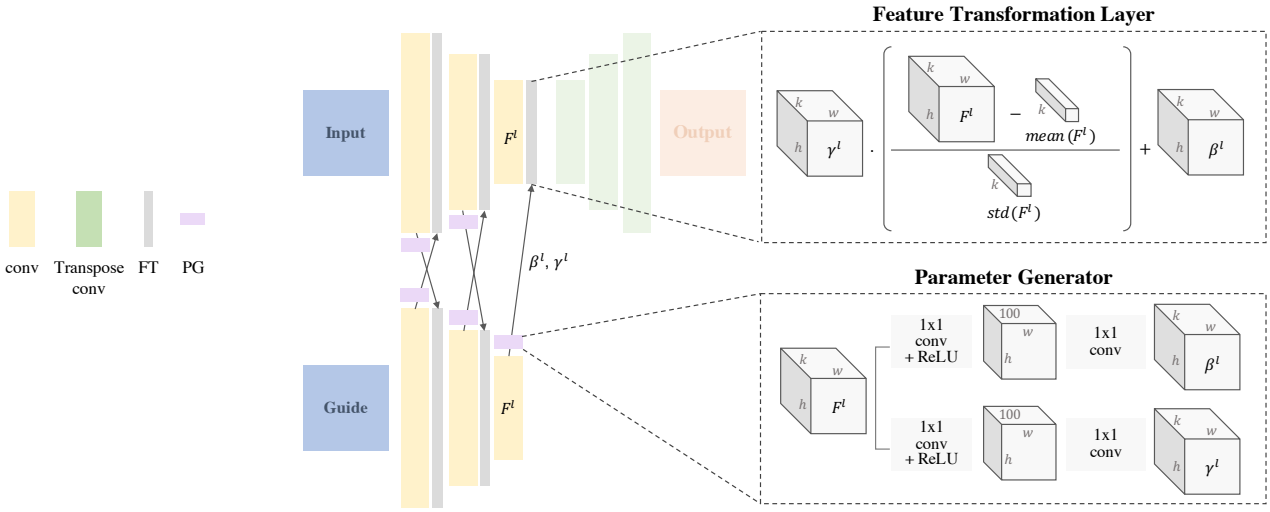
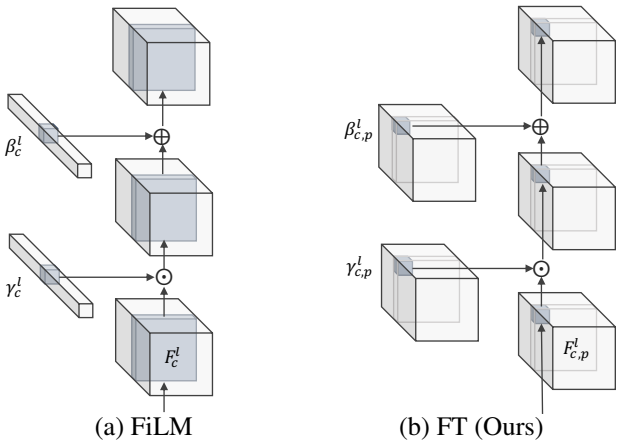


Figure 3. **Bi-directional Feature Transformation.** We present a bi-directional feature transformation model to better utilize the additional guidance for guided image-to-image translation problems. In place of every normalization layer in the encoder, we add our novel FT layer. This layer scales and shifts the normalized feature of that layer as shown in Figure 4. The scaling and shifting parameters are generated using a parameter generation model of two convolution layers with a bottleneck of 100 dimension.



(a) FiLM (b) FT (Ours)

Figure 4. **Feature Transformation (FT).** We present a feature transformation layer to incorporate the guidance into the image-to-image translation model. A key difference between a FiLM layer and our FT layer is that the scaling γ and shifting β parameters of the FiLM layer are *vectors*, while in our FT layer they are *tensors*. Therefore, the scaling and shifting operations are applied in spatially varying manner in our FT layer in contrast to spatially invariant modulation as in the FiLM layer.

3.1. Feature transformation layer

Here, we first present the feature transformation (FT) layer to incorporate the guidance information. In an FT layer, we perform an affine transformation on the normalized input features using scaling and shifting parameters computed from the features of the given guidance image. In Eqn. 1, we show this operation for an l-th layer. The scaling and shifting parameters γ and β are computed from the guidance signal using a *parameter generator* shown in Figure 3.

$$F^{l+1}_{\text{input}} = \gamma^l_{\text{guide}} \frac{F^l_{\text{input}} - \text{mean}(F^l_{\text{input}})}{\text{std}(F^l_{\text{input}})} + \beta^l_{\text{guide}}. \quad (1)$$

A key difference between the FiLM layer [32] and the proposed FT layer is highlighted in Figure 4. Specifically, the scaling γ and shifting β parameters of the FiLM layers are *vectors* and are applied channel-wise. That is, the same affine transformation of feature activations is applied the same way regardless of the spatial position on the feature map. Such approaches are reasonable for tasks such as style transfer or visual reasoning. However, they may not be able to capture fine-grained spatial details that are important for image-to-image translation problems. In contrast, the parameters in our FT layer are three-dimensional *tensors* which offer a flexible way for modulating the input features in a spatially varying manner and supports various forms of guidance signals (e.g., dense, sparse, or multi-channel).

3.2. Bi-directional conditioning scheme

To further utilize the available information from the guidance image, we propose a *bi-directional conditioning scheme*. Unlike existing conditioning schemes that only allow the guidance signal to influence the input image process, our approach supports bi-directional communication between two branches of the networks processing the input and guidance image. This bi-directional flow of information enables the generative model to better capture the

constraints of the guidance image. In our proposed bFT scheme, we replace every normalization layer with our proposed FT layer. At l-th layer, the guidance feature representation manipulates the input feature representation as shown in Eqn. 1, and at the same time is manipulated by that input feature representation. Such that:

$$F^{l+1}_{\text{guide}} = \gamma^{l}_{\text{input}} \frac{F^{l}_{\text{guide}} - \text{mean}(F^{l}_{\text{guide}})}{\text{std}(F^{l}_{\text{guide}})} + \beta^{l}_{input} \qquad (2)$$

Our intuition is that such a bi-directional approach can be seen as a bi-directional communication between a teacher (guidance branch) and a student (input image branch). A one-way communication from the teacher to the student might not help the student understand the teacher as much as two-way communication.

4. Experimental Results

We evaluate our proposed bi-directional feature transformation conditioning scheme on three different guided image-to-image translation problems with three different types of the guidance signal.[1] For all tasks, we use GANs with two possible architectures as our generator model, either Unet or Resnet. We follow the same training objective function (a weighted combination of L_1 loss and an adversarial loss L_{GAN}) as in [17]:

$$L_{GAN}(G,D) + \lambda L_1(G). \qquad (3)$$

where we set λ to 100 for all the experiments. For each task we compare our results with state-of-the-art methods as well as pix2pix [17] (with input concatenation conditioning).

4.1. Controllable sketch-to-photo synthesis

In this texture transfer task, given a sketch and a random sized texture patch as the guidance signal, we aim to synthesize a photo that fills the input sketch respecting that given texture patch.

Implementation details We use the Unet architecture of [17] as the base architecture of our model. For both our bFT model and pix2pix, we train using a learning rate of 0.0002 with 7 layers of Unet architecture. We use an Adam optimizer for both with beta1 as 0.5 for pix2pix, and beta1 as 0.9 for our model. For the handbag dataset, we train for 500 epochs with a batch size of 64. For the shoes and clothes datasets, we train for 100 epochs with batch size of 256.

[1]Code available: https://github.com/vt-vl-lab/Guided-pix2pix

Table 1. Texture Transfer Task: visual quality evaluation using the Learned Perceptual Image Patch Similarity (LPIPS) metric [40] and Frechet Inception Distance (FID) [12] on the datasets generated by [37]. A lower score is better.

	Handbag Dataset		Shoes Dataset		Clothes Dataset	
	LPIPS	FID	LPIPS	FID	LPIPS	FID
Xian *et al*. [37]	0.171	60.848	0.124	44.762	0.113	49.568
pix2pix [17]	0.234	96.31	0.238	197.492	0.439	190.161
Ours	0.161	74.885	0.124	121.241	0.067	58.407

Datasets and metrics We use the 128x128 data generated by Xian *et al*. [37] and follow the same texture patch generation algorithm from the ground truth images. We evaluate the results using the Learned Perceptual Image Patch Similarity (LPIPS) metric proposed by Zhang *et al*. [40] and the frechet inception distance (FID) proposed by Heusel *et al*. [12]. For every sketch in the test set, we generate 10 random sized ground truth texture patches using the texture patch generation algorithm from Xian *et al*. [37] and compute the LPIPS and the FID of the synthesized images. We use the provided pretrained models of Xian *et al*. [37] to compute their results. Their pretrained models are trained on ground truth patches as well as external patches, while our model and pix2pix are trained only on ground truth patches.

Evaluation We show the quantitative results of our work compared to Isola *et al*. [17] and Xian *et al*. [37] in Table 1. While our model training is considerably simpler (trained with only two losses) than that of the Xian *et al*. [37] (with seven different loss terms), we show favorable results against both pix2pix [17] and Xian *et al*. [37] in terms of the LPIPS metric on all three datasets. We also show the FID results.

We show sample qualitative results on the handbag, shoes, and clothes datasets in Figure 5 using ground truth texture patches as the guidance signal.

4.2. Controllable person-image synthesis

In the pose transfer task, given an image of a person and a target pose as a guidance signal, we aim to synthesize an image of that given person in the desired pose.

Implementation details We use ResNet architecture as the base architecture of our model. For both our bFT model and pix2pix, we train for 100 epochs using a learning rate of 0.0002 with a batch size of 8, then we minimize the learning rate to 0.00002 and train for 50 additional epochs. We use the Adam optimizer for both with beta1 as 0.5 for pix2pix, and beta1 as 0.9 for our model. We use 8 layers for the Unet architecture for pix2pix.

Figure 5. **Controllable sketch-to-photo synthesis with texture patches.** Texture transfer qualitative comparison with state-of-the-art-results on the handbags, shoes, and clothes datasets from [37]. Here we use the ground truth texture patches as the guidance signal.

Figure 6. **Controllable person-image synthesis with pose keypoints.** Pose transfer qualitative results on DeepFashion dataset. Our model in general achieves sharper results on this challenging task.

Datasets and metrics We use the 256x256 train and test sets provided by Ma *et al*. [28] from the DeepFashion dataset [26]. Following the evaluation protocols in literature, we use both SSIM and Inception Score (IS) to measure the quality of the synthesized images. We also use the FID metric.

Evaluation We show the quantitative results of our work compared to state-of-the-art methods in Table 2. We note that Siarohin *et al*. [33] trains on a different training set of the DeepFashion dataset and excludes samples where pose keypoints are not detected. To ensure fair comparison, we modify our test set to exclude such samples. We report the results on both the full test set and the modified one. We use the pretrained models provided by [33, 28] to test their models on our test set. We also note that Siarohin *et al*. [33] uses the input pose as an additional input to the model. We show favorable results against other methods using the Frechet Inception Distance (FID).

Note that it is very difficult to measure the quality of a synthesized image. In this task, however, we not only care about the quality of the image, but also about it having the same content and respecting the target pose. We show the qualitative results in Figure 6.

Unlike the aforementioned methods that use keypoint based pose, Neverova *et al*. [30] uses dense pose to per-

Table 2. Pose Transfer task: visual quality evaluation on the DeepFashion dataset [26]. A higher score of SSIM/IS is better. A lower score of FID is better.

	Full test set			Modified test set		
	SSIM	IS	FID	SSIM	IS	FID
Ma *et al*. [29]	0.614	3.29	-	-	-	-
Ma *et al*. [28]	0.762	3.09	47.917	0.764	3.10	47.373
Siarohin *et al*. [33]	0.758	**3.36**	15.655	0.763	**3.32**	15.215
pix2pix [17]	**0.770**	2.96	66.752	**0.774**	2.93	65.907
Ours	0.767	3.22	**12.266**	0.771	3.19	**12.056**

form pose transfer and achieved a score of [SSIM=0.785, IS=3.61], however, we were unable to obtain the data nor the pre-trained model for comparison.

4.3. Depth upsampling

In depth upsampling, we aim to generate a high-resolution depth map given a low resolution depth map with the guidance of a high resolution RGB image.

Implementation details We use the ResNet architecture as the base architecture of our model. For both our bFT model and pix2pix, we only use L1 as the objective function and train for 500 epochs using a learning rate of 0.0002 with batch size of 2. We use an Adam optimizer for both with beta1 as 0.5. For our work, we train on the original size of the data 480x640, however, because pix2pix uses square sized inputs, it is trained on 512x512 resized data and we resize back before evaluation. We use 9 layers for the Unet architecture of pix2pix.

Dataset and metric Following the setting of Li *et al*. [23], we use 1000 samples from the NYU v2 dataset [34] for training and we test on the remaining 449. We generate the low resolution input depth map using bicubic upsampling for three different scale factors 16, 8, and 4. Similar to the works in literature we use RMSE to evaluate the quality of the generated depth.

Evaluation We show the RMSE results of our work compared to Isola *et al*. [17] and state-of-the-art methods in Table 3. We report the results by Li *et al*. [23]. We also show qualitative results for the three scale factors in Figure 7. Our model, while not designed for depth upsampling, can achieve state-of-the-art performance.

4.4. Ablation study

We conduct an ablation study to the effectiveness of our proposed bi-directional conditioning scheme.

Table 3. Depth Upsampling task: root mean square error (RMSE) results in centimeters for the NYU v2 dataset [34].

Depth Scale	$x4$	$x8$	$x16$
Bicubic	8.16	14.22	22.32
MRF [6]	7.84	13.98	22.20
GF [11]	7.32	12.98	22.03
JBU [18]	4.07	13.62	22.03
Ham [10]	5.27	12.31	19.24
DMSG [16]	3.48	6.07	10.27
FBS [1]	4.29	8.94	14.59
DJF [22]	3.54	6.20	10.21
DJFR [23]	3.38	5.86	10.11
pix2pix [17]	4.12	6.48	10.17
Ours	**3.35**	**5.73**	**9.01**

Conditioning schemes We compare our proposed bi-directional feature transformation scheme (bFT) to uni-directional feature transformation (uFT), feature concatenation, and input concatenation schemes shown in Figure 2. We show quantitative results in Table 4.

Number of feature transformation (FT) layers In our bFT model, we use FT in place of every normalization layer. For pose transfer and depth upsampling tasks, we use a Resnet base with 4 normalization layers. Replacing those layers with our proposed FT layer, we end up with 4 FT layers. We compare our approach with using FT at 1, 2, and 3 layers both bi-directionally and uni-directionally. We show the quantitative results in Table 5.

Different approaches to affine transformation Using our bi-directional approach, we compare our proposed FT with CIN and AdaIN. In both CIN and AdaIN, we use FiLM layer in place of every normalization layer. In CIN, we learn the scaling and shifting parameters, while in AdaIN, we use the mean as the scaling parameter and the standard deviation as the shifting parameter. We also test feature transformation at only the last layer of the encoder and compare the performance of our FT with CIN and AdaIN. We show the quantitative results in Table 6.

4.5. User study

We conduct a user study on pair-wise comparisons. We ask 100 subjects to answer 4 random pair-wise comparisons per task and dataset. We ask the subject to select the image that looks more realistic respecting the input and the given guidance signal. We show the user study results in Figure 8.

4.6. Limitation

In the task of texture transfer, we observe a limitation of our work when the guidance patch does not go well with the input sketch. In such a case, the color of the guidance patch

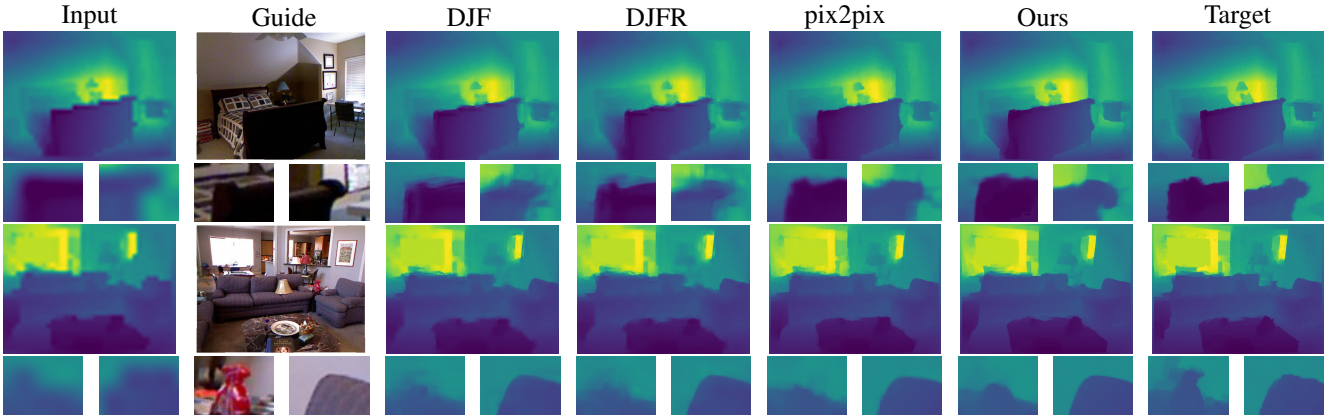


Figure 7. **Depth upsampling guided by an RGB image.** Comparison of depth upsampling qualitative results for a scale factor of 16 with the state-of-the-art methods. The zoomed-in crops show that our method is able to capture fine details with sharper edges.

Table 4. Conditioning schemes.

Conditioning method	Depth Upsampling			Pose Transfer			Texture Transfer					
							Handbags		Shoes		Clothes	
	4x	8x	16x	SSIM	IS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID
Input Concatenation	6.65	8.42	11.86	0.782	3.10	42.330	0.182	85.600	0.137	124.973	0.061	60.795
Feature Concatenation	6.67	7.63	11.59	0.770	3.26	14.672	0.196	87.052	0.145	104.227	0.085	44.900
uFT	5.55	7.26	11.41	0.765	3.18	13.988	0.174	85.273	0.126	119.588	0.071	56.66
bFT (Ours)	3.35	5.73	9.01	0.767	3.17	13.240	0.171	80.179	0.123	119.832	0.067	58.467

Table 5. Number of feature transformation (FT) layers.

#Layers	Depth Upsampling		Pose Transfer					
	uFT	bFT		uFT			bFT	
	x16	x16	SSIM	IS	FID	SSIM	IS	FID
1	10.79	10.79	0.786	2.92	59.678	0.786	2.92	59.678
2	10.75	8.96	0.784	2.98	47.411	0.785	3.01	51.458
3	10.26	8.82	0.768	3.15	16.069	0.766	3.24	13.392
4	11.41	9.01	0.765	3.18	13.988	0.767	3.17	13.240

Table 6. Different approaches to affine transformation.

Method	Depth Upsampling	Pose Transfer		
	x16	SSIM	IS	FID
Ours	9.01	0.767	3.17	13.240
bi-directional AdaIN	13.36	0.722	3.37	160.846
bi-directional CIN	13.97	0.721	3.36	157.335
Final Layer - FT	11.40	0.769	3.25	18.292
Final Layer - AdaIN	14.30	0.720	3.30	146.596
Final Layer - CIN	14.51	0.720	3.58	168.503

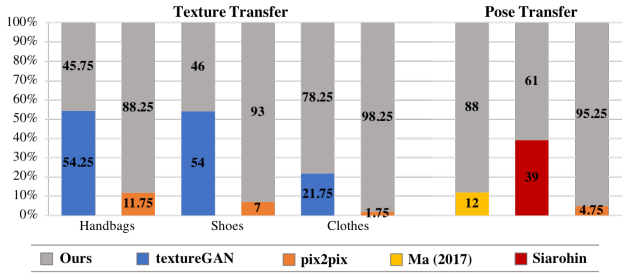


Figure 8. **User Study.** The percentage of people that find our method more realistic respecting the input and guidance signal over state-of-the-art methods using pair-wise comparisons.

Figure 9. **Failure examples.** When the guided patch does not match well with the given sketch, our model fails to hallucinate the given texture.

would propagate through the sketch without fully respecting its texture as shown in Figure 9.

5. Conclusion

We have presented a new conditional scheme for guided image-to-image translation problems. Our core technical contributions lie in the use of *spatially varying* feature transformation and the design of *bi-directional conditioning* scheme that allow the mutual modulation of the guidance and input network branches. We validate the applicability of our method on various tasks. While being application-agnostic, our approach achieves competitive performance with the state-of-the-art. The generality of our method opens promising direction of incorporating a wide variety of constraints for image-to-image translation problems.

Acknowledgment. This work was supported in part by NSF under Grant No. 1755785. We thank the support of NVIDIA Corporation with the GPU donation.

References

[1] Jonathan T Barron and Ben Poole. The fast bilateral solver. In *ECCV*, 2016. 7

[2] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017. 2

[3] Huiwen Chang, Ohad Fried, Yiming Liu, Stephen DiVerdi, and Adam Finkelstein. Palette-based photo recoloring. *ACM Transactions on Graphics (TOG)*, 34(4):139, 2015. 2

[4] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. CrDoCo: Pixel-level domain transfer with cross-domain consistency. In *CVPR*, 2019. 2

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 2

[6] James Diebel and Sebastian Thrun. An application of markov random fields to range sensing. In *NeurIPS*, 2006. 7

[7] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. 2017. 2, 3

[8] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*, 2017. 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, 2014. 2

[10] Bumsub Ham, Minsu Cho, and Jean Ponce. Robust image filtering using joint static and dynamic guidance. In *CVPR*, 2015. 7

[11] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *TPAMI*, 35(6):1397–1409, 2013. 7

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 2

[14] Xun Huang and Serge J Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2, 3

[15] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*. 2

[16] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *ECCV*, 2016. 7

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 1, 2, 5, 6, 7

[18] Johannes Kopf, Michael F Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. In *ACM Transactions on Graphics (ToG)*, volume 26, page 96, 2007. 7

[19] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 3

[20] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 2

[21] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM transactions on graphics*, volume 23, pages 689–694, 2004. 2

[22] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *ECCV*, 2016. 7

[23] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Joint image filtering with deep convolutional networks. *TPAMI*, 2019. 3, 7

[24] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 3

[25] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 2

[26] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 6, 7

[27] Qing Luan, Fang Wen, Daniel Cohen-Or, Lin Liang, Ying-Qing Xu, and Heung-Yeung Shum. Natural image colorization. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 309–320, 2007. 2

[28] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NeurIPS*, 2017. 3, 6, 7

[29] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018. 3, 7

[30] Natalia Neverova, Rıza Alp Güler, and Iasonas Kokkinos. Dense pose transfer. In *CVPR*, 2018. 3, 6

[31] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2

[32] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. 2018. 2, 3, 4

[33] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018. 3, 6, 7

[34] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 7

[35] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. 2018. 2

[36] 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 *CVPR*, 2018. 2

[37] Varun Agrawal Amit Raj Jingwan Lu Chen Fang Fisher Yu James Hays Wenqi Xian, Patsorn Sangkloy. Texturegan: Controlling deep image synthesis with texture patches. *CVPR*, 2018. 3, 5, 6

[38] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 2

[39] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 3

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[41] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics*, 9(4), 2017. 2, 3

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2

[43] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017. 2