Shapes and Context: In-the-Wild Image Synthesis & Manipulation

Aayush Bansal    Yaser Sheikh    Deva Ramanan
Carnegie Mellon University
{aayushb,yaser,deva}@cs.cmu.edu

Figure 1. Our approach synthesizes images from a label mask by non-parametric matching of shapes, parts, and pixels. We show example results for diverse “in-the-wild” scenes containing large amounts of variation in object composition and deformation.

Abstract

We introduce a data-driven model for interactively synthesizing in-the-wild images from semantic label input masks. Our approach is dramatically different from recent work in this space, in that we make use of no learning. Instead, our approach uses simple but classic tools for matching scene context, shapes, and parts to a stored library of exemplars. Though simple, this approach has several notable advantages over recent work: (1) because nothing is learned, it is not limited to specific training data distributions (such as cityscapes, facades, or faces); (2) it can synthesize arbitrarily high-resolution images, limited only by the resolution of the exemplar library; (3) by appropriately composing shapes and parts, it can generate an exponentially large set of viable candidate output images (that can say, be interactively searched by a user). We present results on the diverse COCO dataset, significantly outperforming learning-based approaches on standard image synthesis metrics. Finally, we explore user-interaction and user-controllability, demonstrating that our system can be used as a platform for user-driven content creation.

1. Introduction

We introduce a data-driven model for interactively synthesizing diverse images from semantic label input masks. Specifically, we seek to design a system for in-the-wild image synthesis that is controllable and interpretable. While content creation is a compelling task in of itself (a classic goal of computer graphics), image synthesis is also useful for generating data that can be used to train discriminative visual recognition systems [29]. Synthesized data can be used to explore scenarios that are difficult or too dangerous to sample directly (e.g., training an autonomous perception system on unsafe urban scenes [31]). Figure 1 shows images synthesized using our approach, where the input is a semantic label image.

Parametric vs Nonparametric: Current approaches for image synthesis and editing can be broadly classified into three categories. The first category uses parametric machine learning models. The current state-of-the-art [10, 33, 55] relies on deep neural networks [39] trained with adversarial losses (GANs) [21] or perceptual losses [34] to create images. These approaches work remarkably well when trained on datasets with somewhat limited diversity, such as cityscapes [11], faces [2, 52], or facades [54]. But it is unclear how to extend such approaches for “in-the-wild” image synthesis or editing: parametric models trained on one data distribution (e.g. cityscapes) do not seem to generalize to others (e.g. facades), a problem widely known as dataset bias [53]. The second category of work [1, 14, 15, 27, 38, 48] uses non-parametric nearest neighbors to create content. These approaches have been demonstrated on interactive image editing tasks such as object insertion [38] or scene completion [25]. Though a large inspiration for our own work, such approaches have interestingly fallen out of favor in recent history.

Does more data help? A peculiar property of many parametric synthesis methods is that they do better with less data [2, 21, 33, 45, 58, 62]. The culprit seems to be that such
Figure 2. **Limitations of current approaches for images synthesis:** (a) Current image synthesis models tend to be trained on datasets with somewhat limited diversity, such as cityscapes [11], faces [52], or facades [54]. For example, the average label mask for Cityscapes [11] clearly reveals redundant structure such as a car hood, road, and foliage. In contrast, the average image for COCO [40] is much less structured, suggesting it is a more varied dataset. (b) Indeed, we train state-of-the-art neural architectures [33, 55] on COCO and observe poor convergence (even after a month of training!) resulting in a mode collapsed and averaged outputs. (c) In contrast, our simple matching-based approach is able to synthesize realistic image content by matching to exemplar shapes. In order to generate high-quality images, we find it crucial to encode scene context and part deformation in the matching process - i.e., matching global shapes alone will produce poor images with missing regions due to shape mismatches.

Composition by parts: In this work, we make three observations that influence our final approach; (1) humans can imagine multiple plausible output images given a particular input label mask. We see this rich space of potential outputs as a vital part of the human capacity to imagine and generate. Most parametric networks tend to formulate synthesis as a one-to-one mapping problem, and so struggle to provide diverse outputs (a phenomena also known as mode collapse). Important exceptions include [3, 10, 20, 63] that generated multiple outputs by employing various modifications. (2) visual scenes are exponentially complex due to many possible compositions of constituent objects and parts. It is tempting to combine both observations, and generate multiple outputs by composing scene elements together. But these compositions cannot be arbitrary - one cannot freely swap out a face with a wheel, or place an elephant on a baseball field. To ensure consistency, our matching process makes use of implicit contextual semantics present in the library of exemplar label masks. (3) given an exemplar set of sufficient size, nearest neighbor methods may still perform well with simple features that are not learned (e.g., pixel values). We combine our observations to construct an image synthesis system that exemplar shapes and parts using simple pixel features.

**Our Contributions:** (1) we study the problem of visual content creation and manipulation for in-the-wild settings, and observe that reliance on parametric models lead to averaged or mode-collapsed outputs; (2) we present an approach that utilize shapes and context to generate images consisting of rigid and non-rigid objects in varied backgrounds, and different environmental and illumination conditions; (3) we demonstrate the controllable and interpretable aspects of
our approach that enables a user to influence the generation and select examples from many outputs.

2. Background

Our work combines various ideas on shapes, deformable parts, context, and non-parametric approaches developed in last two decades. We position each separately and the specific insights for their particular usage.

**Shapes & Non-parametric approaches:** Shapes [37, 41, 46] emerge naturally in our world due to its compositional structure. If we had an infinite data-source with all potential shapes for all the objects, then our world could be represented by a linear combination of different shapes [23, 46]. In this work, we aim to generate images from semantic and instance label maps as input. Meaningful shapes and contours [5, 24] makes for an obvious interpretation for such input. In an unconstrained in-the-wild data distribution consisting of both rigid and non-rigid objects, it become non-trivial to model such shapes for a one-to-many mappings. We, therefore, want to leverage the shape information explicitly from the training data in our formulation by simple *copy-pasting*. Non-parametric methods [14, 15, 16, 19, 27, 35] find their use for various computer vision tasks such as texture-synthesis [15, 16], image super-resolution [19], action recognition [14], or scene completion [25].

Our work draws similarity to idea of scene composites [32, 44, 49, 48]. Russell et al [48] use shapes or scene composites to query matches for semantic segmentation using LabelMe dataset [50]. In an another work, Russell et al [49] use similar idea of composites for object discovery. Isola and Liu [32] use this idea of composites for scene parsing and collaging. Recently, Qi et al [44] used shapes in a semi-parametric form to synthesize images from a semantic label map. These different approaches [44, 48] on scene composites or shapes are however constrained to rigid and non-deformable objects from a constrained data-distribution such as road-side scenarios from LabelMe [50], or Cityscapes [11]. Our work extends the prior work to non-rigid and deformable shapes from an unconstrained in-the-wild data distribution. Figure 2-(c) shows how global shapes are insufficient and one needs to consider local information about parts and pixels.

**Deformable Objects & Parts:** The global shape fitting can be reliably estimated for non-deformable objects but local shapes or *parts* [7, 18, 22] are required when considering non-rigid objects. The prior work on local components [7], regions [22], or parts [18, 51, 57] has largely been focused on recognition. On the other hand, our work draws insight from ideas on compositional matching [8, 17] and we use the parts, components, and regions for synthesizing images. In this work, we generate parts from various global shapes to do image synthesis. This enables us to consider local information without any explicit part labels.

**Context as a major cue:** Context is a natural and powerful tool to put things in perspective [6, 30]. There is a wide literature on the use of context in computer vision community [13, 42] and is beyond the scope of this work to illustrate them completely. In this work, we use contextual information at both global and local level to do better and faster matching of global shapes, parts, and pixels. The contextual information, while itself remaining in background, enables us to do an effective non-parametric matching.

**User Controllable Content Synthesis & Manipulation:** Multiple works [3, 4, 38, 61, 59, 55] in computer graphics and vision literature have demonstrated the importance of user-controlled image operations. Grab-cut [47] enables user-based segmentation of a given scene. Lalonde et al [38] use a non-parametric approach to insert objects in a given image. Kholgade et al [36] demonstrate a user-controlled 3D object manipulation. In this work, we demonstrate how shapes can be used naturally and intuitively for a user controllable content creation and manipulation.

3. Method

Given a semantic and an instance label map, $X$, our goal is to synthesize a new image, $Y$. Our formulation is a hierarchical non-parametric matching ensuring the following stages in order: (1) global scene context; (2) instance shape consistency; (3) local part consistency; and finally (4) minute pixel-level consistency.

**Global Scene Context:** In a big data settings with hundred thousands and million examples, doing nearest neighbors could be a time-consuming process. We make this process faster by using global scene context to prune the list of training examples from which the shapes should be extracted. Only those examples are considered if they fall in one of three categories: (1) their global image has same labels as input; (2) the labels in input is its subset; (3) the labels in input is its superset. This reduces the search space from hundred thousand shapes to a few hundreds. We further prune them to top-$N$ images for searching shapes by computing a global coverage and a pixel coverage score.

A **global coverage** score is computed to ensure the top-$N$ label maps in the training set have similar distribution of labels as are in a given query label map. We compute the normalized histogram of labels (both query and training), and compute a $l_2$ distance between query and training label map. A **pixel coverage** score is computed to ensure we select the images with maximum pixel-to-pixel overlap. This score is computed by aligning a query label map and an example from training set, followed by the hamming distance between them. To make it faster, we resize the images to $100\times100$ and then compute the normalized hamming distance between the respective labels. We sum both global coverage and pixel coverage scores, and choose $N$
Shape Consistency: We seek shapes as the first step to define different components in an image. We represent the shapes in an instance and semantic label mask as \( \{x_1, x_2, \ldots, x_N\} \) where \( N \) is the total number of shapes for a given input. Each shape has an associated semantic label \( l : l \in \{1, 2, \ldots, L\} \) where \( L \) is the number of unique labels. We then make a tight bounding box around this shape \( x_i \) so that it could be used as a rectangular convolutional filter \( (w_i) \) to retrieve similar shapes from the training data. We represent this filter using: (1) a simple logical operator: the part of a shape \( (x_i) \) in the filter \( (w_i) \) is set to 1, and the remaining part is set to \(-1\). This forces the filter to search the composites with boundaries and details; and (2) a contextual operator: we extract the labels from the input label mask for this filter. This information is to force our matching function to extract the shapes which have similar context.

We use the logical operator \( (w^l) \) and contextual operator \( (w^c) \) to rank the remaining shapes for our query component using the scoring function:

\[
S_{\text{shape}}(w_i, w_j) = w^l_i * w^l_j + \sum_{k=1}^{N_s} I(w^c_{i,k} - w^c_{j,k}), \quad (1)
\]

where \( I \) is an indicator function, and \( N_s \) is the total number of pixels in a given query component. Since we have fixed the size to \( 50 \times 50 \) in our formulation, \( N_s = 2500 \). We use this scoring function (Eq. 1) to score different shapes in our pruned list for a given shape. The RGB component for an extracted shape is its intersection with the query shape, i.e. only pixels active in both extracted and query shape are considered. Figure 3-left shows the part of our algorithm to compute shape-consistency score.

We ignore the shapes if the ratio of their aspect-ratio to that of query component is either less than 0.5 or greater than 2. Finally, we make this convolution processing faster by using a fixed size filters and low-res label masks of \( 50 \times 50 \). This also helps us to generate composite of arbitrarily high resolution without any extra computational cost.

Part Consistency: Modeling occlusions and deformable aspects of non-rigid objects in real world are extremely hard. The problem is even aggravated with noisy shape inputs. The insufficient shape data and non-rigid objects in real world leads to parts and local regions [7]. We seek parts from top-\( k \) global shapes. Importantly, the part in-
Non-parametric matching: Our approach for generating images from a label mask by non-parametric matching of global shapes, local parts, and pixel-consistency. The above examples contain varying background, cluttered environment, varying weather and illumination conditions, and multiple rigid and non-rigid objects in various shapes and forms.

formation is required when a global shape is not able to capture. We extract the knowledge of parts from the global shapes in a spirit similar to non-parametric texture synthesis [16]. The shape components are resized to 256×256, so that local information can be well searched. We extract a 16×16 patches from the resized global shape template. Local contextual information (similar to HOG [12], or Group-Normalization [56]) is used by considering the neighboring 8 patches. The parts are scored using:

$$S_{part}(w^p_i, w^p_j) = \sum_{k=1}^{N_p} I(w^p_{i,k} - w^p_{j,k})$$

where $I$ is an indicator function, and each patch ($w^p$), is represented by a $N_p$ (256×9) dimensional vector containing the label information in the patch. Importantly, we do not need to look in a larger window for part matching as we have weakly aligned global shapes. Therefore, we restrict the patches to look in a surrounding 5×5 patch window. This corresponds to 112×112 pixel window in a resized global shape template. To copy the RGB component, we take an average of top-3 retrieved patch windows. Figure 3-middle shows the part of our algorithm to compute part-consistency score.

Pixel Consistency: The shapes and parts have accounted for most of the non-parametric image synthesis. However, they does not ensure pixel-level consistency and often ends up with minor holes in an image. We enforce a pixel-level consistency in this process to account for the remaining holes in synthesized image. This process is similar to our part consistency algorithm, except that it is done on every pixel. Each pixel is represented by a surrounding 11×11 window. We use the criterion in Eq. 2 to compute similarity between two feature vectors. To expedite this matching, we compute features for a low-res input label map (128×128) as pixel consistency is ensured to fill minor holes alone. Finally, we look in surrounding region of 5×5 from a 128×128 image to fill in the information as global and local consistency have already been accounted by shape and part consistency.

Hierarchical Composition: We combine the information hierarchically from shapes, parts, and pixels to generate a full image. Figure 3-right shows the composition starting from an input label mask. Firstly, we use the global shape component to fill the major chunk of images. The missing information is then filled using the local part consistency. Finally, the minor holes are filled using the pixel-level consistency. The combination of these three stages enable us to generate a image from an input label mask by simple non-parametric matching. Figure 4 shows outputs generated by our approach for varying background, cluttered environment, varying weather and illumination conditions, and multiple rigid and non-rigid objects in various
shapes and forms.

**Generating Multiple-Outputs:** A salient aspect of considering shapes and parts in a non-parametric matching provides multiple outputs for free. We use the extracted shapes and parts, and can combine them in exponential ways without any extra overhead. We show multiple examples synthesized for a given label mask using our approach in Figure 6. Generating these multiple outputs is not trivial when using parametric approaches [10, 33], and a substantial research [20, 63] has been conducted for this process. However, it is just trivial for non-parametric matching.

**User Controllable Content Creation:** We finally demonstrate the applicability of our approach for a user controllable content creation in Figure 7. Note how our approach can be easily used to edit the label mask by inserting shapes to generate a new output. More importantly, synthesis and manipulation aspects go hand-in-hand for our approach. A human user can clearly interpret and influence any stage of synthesis, and can easily generate a different output by varying a shape. Manipulation naturally emerges in our non-parametric approach without any additional efforts. This is not true for prior parametric approaches that require a specialized machinery for this task.

---

4. Experiments

**Dataset:** We use semantic and instance label mask from COCO [40] to study the problem of in-the-wild image synthesis and manipulation. This dataset consists of 134 different objects and stuff categories making it the most diverse and varied publicly available dataset. There are 118,287 images in the training set (40× more than cityscapes [11]), and 5,000 images in the validation set (100× more than cityscapes). We use the paired data of labels and images from training set to extract global shapes and synthesize parts and pixels. The images are synthesized using semantic and instance label masks in validation set. Our approach does not require any training, and therefore can use the labels and image component from anywhere. For the sake of fair comparison with parametric approaches, we restrict ourselves to COCO training data.

**Baselines:** To the best of our knowledge, there does not exist a non-parametric approach that has attempted the problem of the in-the-wild image synthesis from label masks. We, therefore, compare our approach with parametric approaches: (1) Pix2Pix [33]; and (2) Pix2Pix-HD [55], using their publicly available codes. The complexity, diversity, and size of this dataset makes it a computational challenge for a generative parametric approach to deal with. Training a simple Pix2Pix model took 20 days on a single
Nvidia Titan-X GPU. On the same compute, we trained a Pix2Pix-HD model for a month but did not observe any convergence. It may be possible that a reasonable Pix2Pix-HD model be trained if we let the training go longer for an extra month or two, or use advanced computational resources. It may also be due to design of architecture and hyper-parameters specifically suited for Cityscapes, and that efforts are required to tune hyper-parameters to make it work for a large and diverse dataset as COCO. For the sake of fair comparison, we additionally use Cityscapes to contrast our approach with prior works [10, 33, 44, 55] even when it comes at the cost of performance for our own approach due to limited data. Additionally, we resize our generated outputs to 256 × 256 just to make a fair comparison with Pix2Pix on COCO. However, we can generate outputs having same resolution as that of input label masks without any increase in compute.

**FID Scores**: We compute FID scores [28] using the images generated from different approaches. Lower FID values suggest more realism. Table 1 contrast FID scores computed on generated images (COCO) with Pix2Pix and Pix2Pix-HD (resized to 256 × 256 and 64 × 64 resolution). Without using any oracle, the top-1 example generated from our approach significantly outperforms the prior work. Additionally, note the performance improvement due to each stage in our hierarchical composition.

**Mask-RCNN Scores**: We use a pre-trained Mask-RCNN [26] to study the quality of synthesis on COCO [40] for Pix2Pix and our approach. This model is trained for 80 object categories of COCO dataset. While it is trained for instance segmentation, we use its output and convert it to semantic labels for consistency in evaluation. Our goal is to observe if we can get the same class labels from the synthesized images as one would expect from a real image. We, therefore, run it on original images from the validation set and use these pseudo semantic labels as ground truth for evaluation. Next, we run it on synthesized images and contrast it with the labels from original image. To measure the performance, we use three criterion: (1) mean pixel accuracy (PC); (2) mean class accuracy (AC); (3) mean intersection over union (IoU). Higher the score for each of the criterion, better is the quality of synthesis. Table 2 contrasts the performance of our approach with Pix2Pix and demonstrates substantially better results. Our performance improves when an oracle is used to select the best from five outputs. Note that top-100 examples from the training set are used for global shape matching in our approach for this experiment.

**Human Studies**: We did human studies on a randomly selected 500 images. We show the outputs of Pix2Pix, Pix2Pix-HD, and our approach (randomly picked one output from multiple) to human subjects for as much time as they need to make a decision. We asked them to pick one that looks close to a real image. The users were advised to use ‘none of these’ if all approaches are consistently bad. 51.2% times user picked an output generated from our approach, 7.8% times the outputs from Pix2Pix, and preferred ‘none of these’ 41% times. The human studies suggest that while our approach is most likable but there are still many
Figure 7. User-Intervention & Image Manipulation: The first two columns show an original label mask and a synthesized output for it using our approach. A user can add a shape to the label mask, and a new output will be generated by matching the corresponding shape.

<table>
<thead>
<tr>
<th>Method</th>
<th>#examples</th>
<th>Oracle</th>
<th>PC</th>
<th>AC</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pix2Pix [33]</td>
<td>1</td>
<td>x</td>
<td>17.9</td>
<td>8.9</td>
<td>4.9</td>
</tr>
<tr>
<td>Non-Parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>1</td>
<td>x</td>
<td>44.5</td>
<td>31.0</td>
<td>20.9</td>
</tr>
<tr>
<td>Ours</td>
<td>5</td>
<td>✓</td>
<td>58.2</td>
<td>41.2</td>
<td>31.4</td>
</tr>
</tbody>
</table>

Table 2. Mask-RCNN Scores on COCO: We use a pre-trained Mask-RCNN model [26] to study the quality of image synthesis. We run it on synthesized images and contrast it with the labels from original image. To measure the performance, we use three criterion: (1) mean pixel accuracy (PC); (2) mean class accuracy (AC); (3) mean intersection over union (IoU). Higher the score for each of the criterion, better is the quality of synthesis. We outperform Pix2Pix. The performance further improves significantly when an oracle is used to select from five examples.

Cityscapes: Table 3 contrasts the performance of our approach with prior approaches [10, 33, 44, 55] that have specifically demonstrated on Cityscapes. Except Pix2Pix, we used publicly available results for this evaluation. Our approach performs second to Pix2Pix-HD and better than prior parametric and semi-parametric approaches with just 25 images to extract shapes and parts to compose a new image from semantic labels. The performance improves when using an oracle to select best amongst the 5 generated outputs. Our performance may further improve as we increase the number of global images to do shape and part extraction.

<table>
<thead>
<tr>
<th>Method</th>
<th>#examples</th>
<th>Oracle</th>
<th>PC</th>
<th>AC</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pix2Pix [33]</td>
<td>1</td>
<td>x</td>
<td>72.5</td>
<td>29.5</td>
<td>24.6</td>
</tr>
<tr>
<td>CRN [10]</td>
<td>1</td>
<td>x</td>
<td>49.0</td>
<td>22.5</td>
<td>18.2</td>
</tr>
<tr>
<td>Pix2Pix-HD [55]</td>
<td>1</td>
<td>x</td>
<td>79.0</td>
<td>43.3</td>
<td>37.8</td>
</tr>
<tr>
<td>Semi-Parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIMS [44]</td>
<td>1</td>
<td>x</td>
<td>68.6</td>
<td>35.1</td>
<td>28.1</td>
</tr>
<tr>
<td>Non-Parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours (top-25)</td>
<td>1</td>
<td>x</td>
<td>67.1</td>
<td>38.0</td>
<td>30.5</td>
</tr>
<tr>
<td>Ours (top-25)</td>
<td>5</td>
<td>✓</td>
<td>71.3</td>
<td>39.6</td>
<td>32.4</td>
</tr>
</tbody>
</table>

Table 3. PSP-Net Scores on Cityscapes: We use a pre-trained PSP-Net model [60] to evaluate the quality of synthesized images (1024 \times 2048). This model is trained for semantic segmentation on cityscapes. We run the synthesized images through this model, and generate a semantic label map for each image. The semantic label map from the synthesized images is contrasted with the semantic label map from the original image. We compute three statistics for each approach: (1) Mean Pixel Accuracy (PC); (2) Mean Class Accuracy (AC); (3) Mean intersection over union (IoU). For each of these criterion- higher the score, better is the quality of synthesis. With just 25 global nearest neighbors to extract shapes and parts, our non-parametric approach is competitive to the parametric models and semi-parametric models.

5. Discussion & Future Work

We present an exceedingly simple non-parametric approach for image synthesis and manipulation in-the-wild. While the diverse data-distribution and large datasets make it challenging for parametric approaches to operate on, it enables simple matching of shapes and parts to work well. The non-parametric matching enables us to generate exponentially large number of outputs by varying shapes and parts. Importantly, shapes and parts are intuitive to a normal human user as well. This makes our approach interpretable and suitable for user-controllable content creation and editing. The future work in this direction may address smarter ways of combining shapes and parts information, and explore spatiotemporal consistency to do in-the-wild video synthesis and manipulation.
References


