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

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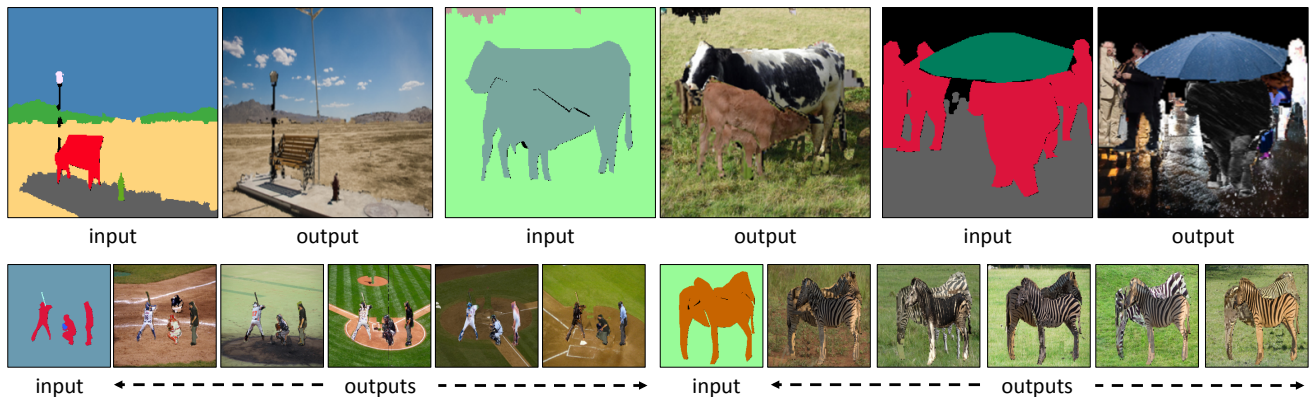Carnegie Mellon University
{aayushb,yaser,deva}@cs.cmu.edu

Figure 1. Our approach synthesize images from label maps 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. We also show multiple outputs generated from a label map using our approach in the bottom row.

Abstract

We introduce a data-driven approach for interactively synthesizing in-the-wild images from semantic label maps. 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 approach for interactively synthesizing diverse images from semantic label maps.

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 [28]. 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 [30]). Figure 1 shows images synthesized using our approach, where the input is a semantic label map.

Parametric vs Nonparametric: Current approaches for image synthesis and editing can be broadly classified into two categories. The first category uses parametric machine learning models. The current state-of-the-art [1, 9, 32, 54] relies on deep neural networks [38] trained with adversarial losses (GANs) [20] or perceptual losses [33] to create images. These approaches work remarkably well when trained on datasets with somewhat limited diversity, such as Cityscapes [10], faces [51], or facades [53]. 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 [52]. The second category of work [13, 14, 26, 37, 47] uses non-parametric nearest neighbors to create content.
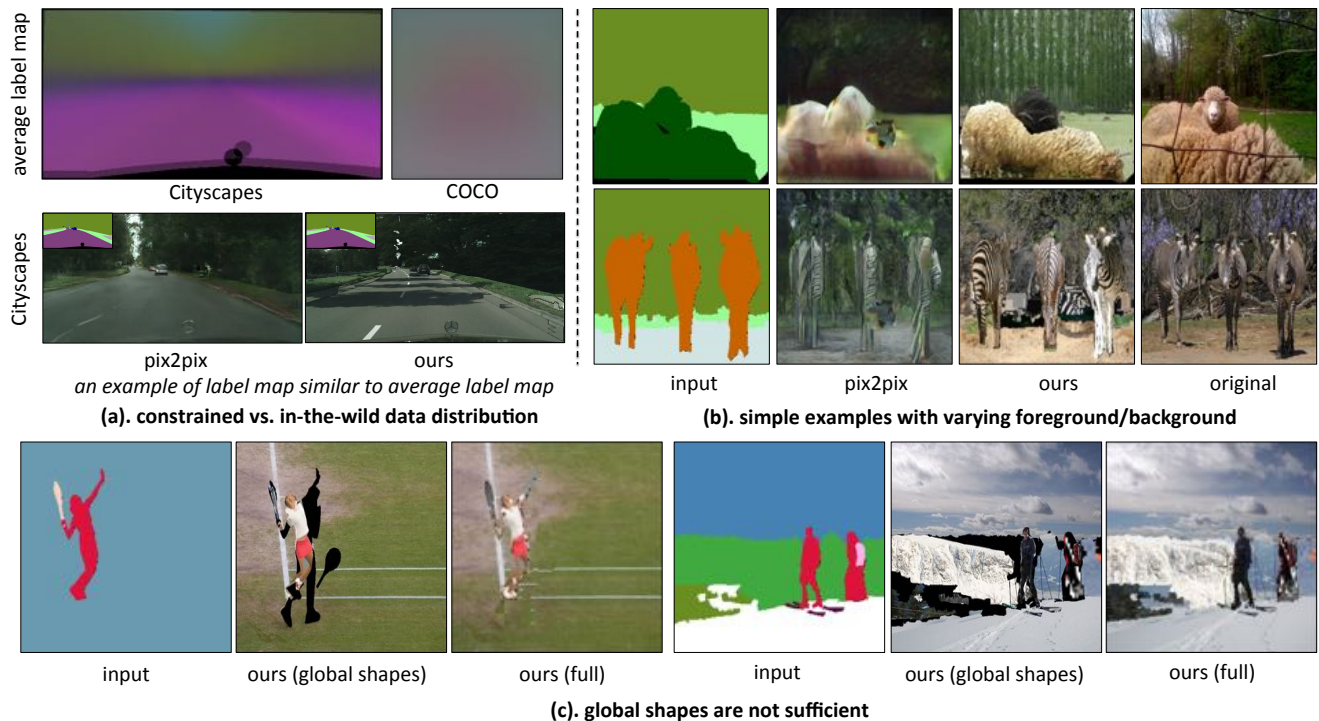
(a). constrained vs. in-the-wild data distribution

average label map

Cityscapes   COCO

Cityscapes

pix2pix   ours

an example of label map similar to average label map

(b). simple examples with varying foreground/background

input   pix2pix   ours   original

(c). global shapes are not sufficient

input   ours (global shapes)   ours (full)

input   ours (global shapes)   ours (full)

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 [10], faces [51], or facades [53]. For example, the average label mask for Cityscapes [10] clearly reveals redundant structure such as a car hood, road, and foliage. In contrast, the average image for COCO [39] is much less structured, suggesting it is a more varied dataset. **(b)** Indeed, we train state-of-the-art neural architectures [32, 54] 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.

These approaches have been demonstrated on interactive image editing tasks such as object insertion [37] or scene completion [24]. Though a large inspiration [26] 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 [1, 20, 32, 44, 57, 61]. The culprit seems to be that such methods don't do well on diverse training sets, and in practice larger training sets tend to be diverse. This is in contrast with truly non-parametric methods that do better with *more* data [24]. Figure 2-(a)-(b) highlights the differences between limited and diverse datasets, using illustrative examples of Cityscapes [10] and COCO [39]. While parametric methods do well on limited data distributions, they struggle to perform on diverse datasets. Recent works [8, 42] have attempted to overcome this challenge by using enormously large model sizes and massive compute (see concurrent work from Park et al [42] for more discussion on parametric approaches).

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 [2, 9, 19, 62] 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 a 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
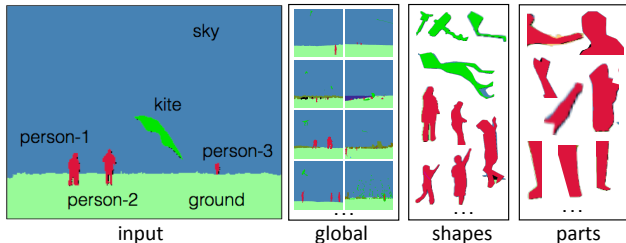
Figure 3. **Overview:** Given a label map (left), our goal is to generate multiple plausible images. We make use of hierarchical matching for efficient retrieval of images, shapes, and parts
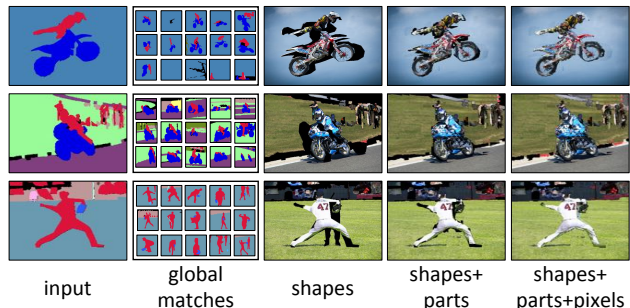

Figure 5. **Hierarchical Composition:** Given an input label map (left), we show the outputs of four stages of our non-parametric matching approach. The second column shows the top-15 global matches for the input. The third column shows the output of composition of global shapes extracted using our shape consistency algorithm. The fourth column shows the improved output by introducing local part consistency to previous output. Finally, minute pixel level holes are filled by our pixel-consistency algorithm.

are not learned (e.g., pixel values). We combine our observations to construct an image synthesis system that relies on 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: Shapes [36, 40, 45] 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 [22, 45]. In this work, we aim to generate images from semantic and instance label maps as input. Meaningful shapes and contours [4, 23] makes for an obvious interpretation for such an input.

Non-parametric approaches: In an unconstrained in-the-wild data distribution consisting of both rigid and non-rigid objects, it becomes 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 [13, 14, 15, 18, 26, 34] find their use for various computer vision tasks such as texture-synthesis [14, 15], image super-resolution [18], action recognition [13], or scene completion [24].

Scene Composites: Our work draws similarity to idea of scene composites [31, 43, 48, 47]. Russell et al [47] use shapes or scene composites to query matches for seman-

tic segmentation using LabelMe dataset [49]. In an another work, Russell et al [48] use similar idea of composites for object discovery. Isola and Liu [31] use this idea of composites for scene parsing and collaging. Recently, Qi et al [43] used shapes in a semi-parametric form to synthesize images from a semantic label map. These different approaches [43, 47] 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 [49], or Cityscapes [10]. 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* [6, 17, 21] are required when considering non-rigid objects. The prior work on local components [6], regions [21], or parts [17, 50, 56] has largely been focused on recognition. On the other hand, our work draws insight from ideas on compositional matching [7, 16] and we use the parts, components, and regions for synthesizing images. In this work, we relax strict shape matching to do local part generation from various global shapes to do image synthesis. This enables us to consider local information without any explicit part labels.

Context: Context is a natural and powerful tool to put things in perspective [5, 29]. There is a wide literature on the use of context in computer vision community [12, 41] and is beyond the scope of this work to illustrate them completely. In this work, the contextual information comes for free with our input, i.e. semantic label map. We use this context at both global and local level to do better and faster matching of global shapes, parts, and pixels. The contextual
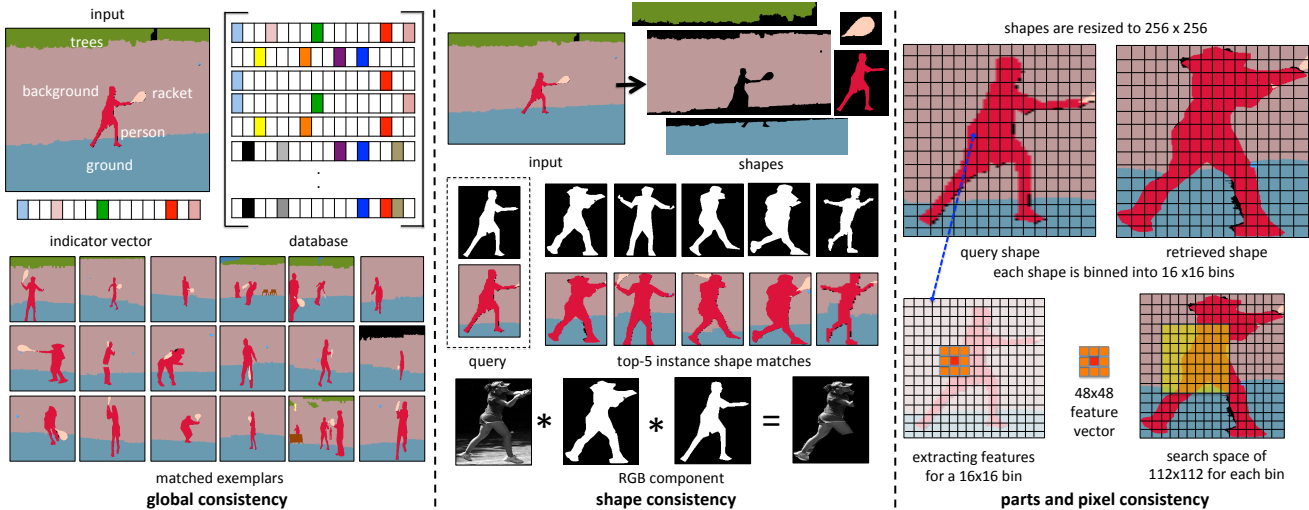
Figure 4. **Four stages of non-parametric matching:** (1) **Global Consistency** - Given a semantic label map, we using an indicator vector that tells what categories are present. This indicator vector helps us to quickly find the relevant examples from database. This prunes away 99% of exemplar database to get required shapes. (2) **Shape Consistency** - We extract various shapes from the input label mask, and match query shape to shapes within the exemplar set by using a shape-and-context feature. We show examples of top-5 retrieved shapes for a query shape on left. The image information from the retrieved shapes is then extracted by considering the mask of query shape and retrieved shape; (3) **Part Consistency** - We observe that global shape retrieved in last stage is missing information about the hands and legs of the query shape (human in this case). We define a local shape matching approach that looks in the neighborhood to synthesize parts. The query and top-k shapes are resized to 256×256, and binned into 16×16 bins with each bin being a 16×16 patch. Each patch is represented by label information contained in it, and an additional 8 neighboring patches. This provides contextual information about the surroundings. The parts are looked in an adjacent 112×112 region and the ones with minimum hamming distance is considered. **Pixel Consistency** algorithm follows a similar setup. See section 3 for more details.

information, while itself remaining in background, enables us to do an effective non-parametric matching.

User Control: Multiple works [2, 3, 37, 54, 58, 60] in computer graphics and vision literature have demonstrated the importance of user-controlled image operations. Grabcut [46] enables user-based segmentation of a given scene. Lalonde et al [37] use a non-parametric approach to insert objects in a given image. Kholgade et al [35] 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 consistency; (2) instance shape consistency; (3) local part consistency; and finally (4) minute pixel-level consistency. Figure 3 gives an overview of our approach.

Global Scene Consistency: In a big data setting with hundred thousands and million examples, doing nearest neighbors could be a time-consuming process. We make this

process faster by using a global indicator vector (or 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 100×100 and then compute the normalized hamming distance between the respective labels. We sum both global coverage and pixel coverage scores, and choose N images in the training set with the lowest scores. This use of global scene context drastically reduces the search space for our

Figure 6. **Non-parametric matching:** Our approach for generating images from a semantic label map 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.

non-parametric approach, and enables to do synthesis on everyday computing devices (single core CPUs instead of GPUs). Figure 4-(left) shows similar examples retrieved by simple matching of indicator vector.

Instance 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 $\{x_1, x_2, ..., x_N\}$ where N is the total number of shapes for a given input. Each shape has an associated semantic label $l : l \in \{1, 2, ..., L\}$ where L is the number of unique labels. We then make a tight bounding box around a shape x_i so that it could be used as a rectangular convolutional filter (w_i) to retrieve similar shapes within the exemplar set. The filter is fixed to 50×50 in this work. We represent a filter/shape using: (1) a simple **mask 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 enables to match the shapes with similar boundaries and details; and (2) a **contextual operator**: we extract the labels from the input label mask for this filter. This information helps in getting better matches with similar context.

We use the mask operator (m, where $m_i[k] \in \{-1, 1\}$) and contextual operator (v, where $v_i[k] \in \{1, 2, ..., N_c\}$ and N_c is number of categories) to match a query shape with shapes in exemplar set using the scoring function:

$$S_{shape}\Big((m_i, v_i), (m_j, v_j)\Big) = \sum_t m_i[t].m_j[t] + I(v_i[t] = v_j[t]), \quad (1)$$

where I is an indicator function with value 1 when true, and zero otherwise. We ignore the shapes from exemplar set when the ratio of their aspect-ratio to that of query shape is either less than 0.5 or greater than 2. Using a fixed size filters and low-res label masks help us to generate composite of arbitrarily high resolution without any extra computational cost. 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 4-(middle) illustrates this part of our algorithm.

Local 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 [6]. We seek parts from top-k global shapes. Importantly, the part information 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 [15]. The shape components are resized to 256×256, so that local information can be well searched. We extract a 16×16 patch from the resized global shape template. Lo-

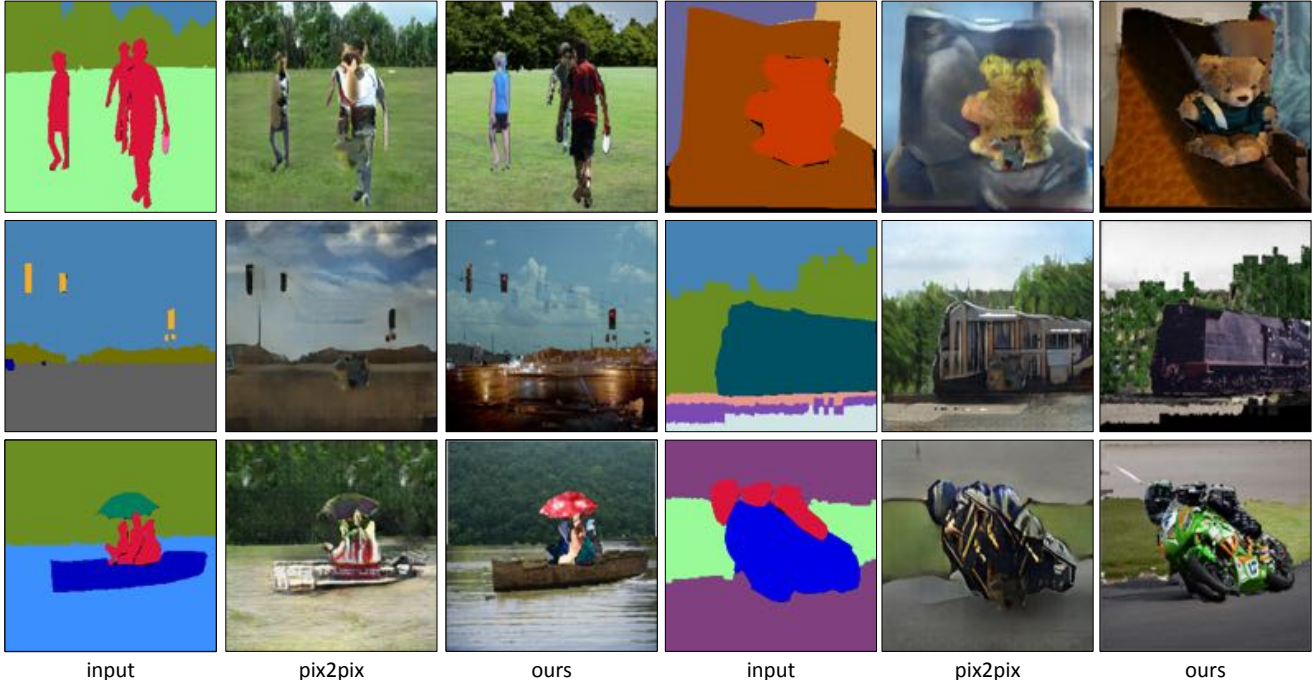| input | pix2pix | ours | input | pix2pix | ours |

Figure 7. **Parametric vs. Non-parametric:** We generate images from an input label map. The second column shows the output of Pix2Pix [32] trained on COCO training set. The third column shows the output of our non-parametric hierarchical matching approach.

cal contextual information (similar to HOG [11], or Group-Normalization [55]) is used by considering the neighboring 8 patches. A patch (p) is, therefore, represented by a 256×9 dimensional vector containing the label information in its region and surroundings. The parts are scored using:

$$S_{part}(p_i, p_j) = \sum_t I(p_i[t] = p_j[t]) \qquad (2)$$

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. We copy the RGB component from the best matching patch window. Figure 4-(right) shows the part of our algorithm to compute part-consistency.

Minute 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 la-

bel 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 5 shows the composition starting from an input label map. Firstly, we find relevant examples using global scene context. We then use the instance shape components to fill the major chunk of image. 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 four stages enable us to efficiently generate a image from an input label mask by simple non-parametric matching.

Qualitative Results: Figure 6 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. Figure 7 contrasts our approach with Pix2Pix [32].

Multiple-Outputs: A salient aspect of considering shapes and parts in a non-parametric matching provides multiple outputs for free. We can combine various extracted shapes in exponential ways without any extra overhead. We show multiple examples synthesized for a given label map using our approach in Figure 8. Generating these multiple outputs is not trivial when using parametric approaches [9, 32], and
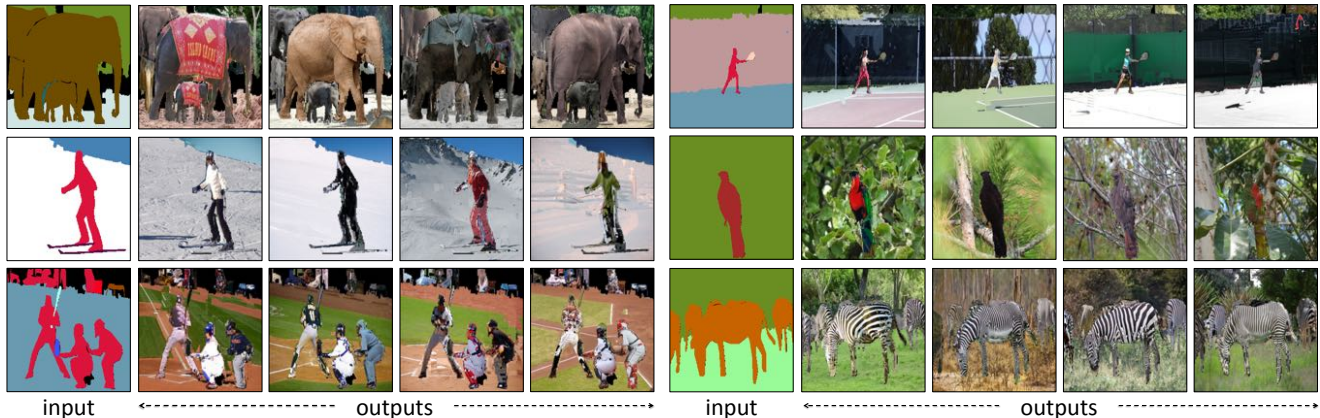
Figure 8. **Multiple Outputs:** Our approach can easily generate exponentially large number of outputs by changing shapes and parts. We show four outputs generated for each label map.

has been a subject of multiple studies [19, 62].

User Control: We finally demonstrate the applicability of our approach for a user controllable content creation in Figure 9. Even though it is rare to see *an elephant on the baseball field*, our approach can easily generate such an example by inserting shapes. 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.

4. Experiments

Dataset: We use semantic and instance label mask from COCO [39] to study the problem of in-the-wild image synthesis and manipulation. This dataset consists of 134 different objects and stuff categories making it one of the most diverse and varied publicly available dataset. There are $118,287$ images in the training set ($40\times$ more than Cityscapes [10]), and $5,000$ images in the validation set ($100\times$ 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 maps on this large scale. We, therefore, compare our approach with parametric approaches: (1). Pix2Pix [32]; and (2). Pix2Pix-HD [54], 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 vanilla 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. It is for this reason we also use Cityscapes to contrast our approach with prior works [9, 32, 43, 54] for the sake of fair comparison. 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 [27] 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 [25] to study the quality of synthesis on COCO [39] 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
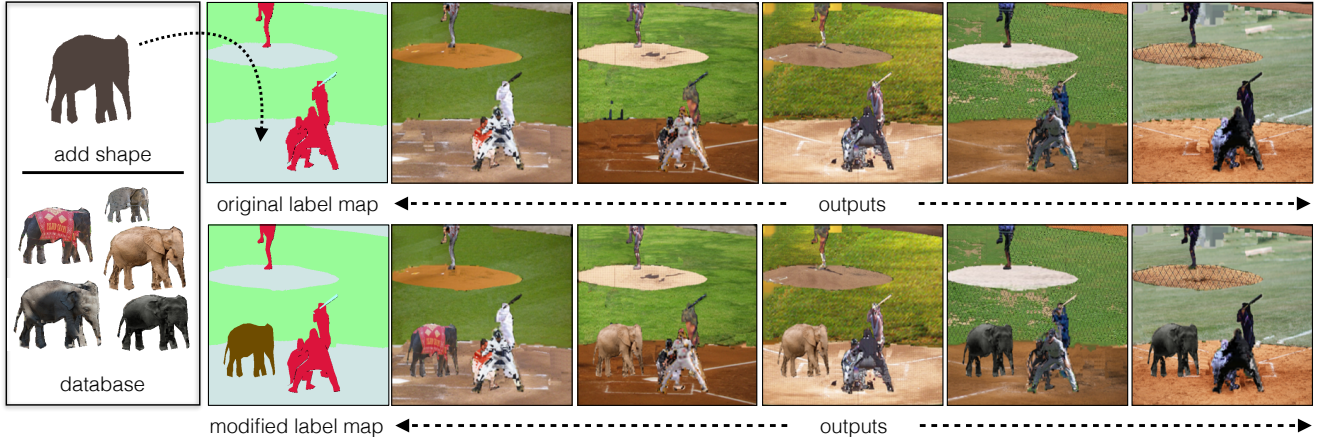
Figure 9. **User-Intervention & Manipulation:** The first row shows multiple outputs generated for a given label map. We insert the shape of *elephant* in this input label map to synthesize an *elephant on the baseball field*. The second row shows multiple outputs for the modified label map.

Method	#examples	Oracle	FID score (256×256)	FID score (64×64)
Pix2Pix [32]	1	✗	70.43	41.45
Pix2Pix-HD [54]	1	✗	157.13	109.49
Ours (shapes)	1	✗	37.26	23.22
Ours (shapes+parts)	1	✗	32.62	18.02
Ours (shapes+parts+pixels)	1	✗	**31.63**	**16.61**

Table 1. **FID Scores on COCO:** We compute FID score [27] to contrast the realism in outputs produced by different approaches. **Lower FID values suggest more realism**. We observe that our approach outperforms prior approaches significantly. We also demonstrate as how different stages in our hierarchical composition leads to better outputs.

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 exemplar matches are used for instance shape matching in 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

Method	#examples	Oracle	PC	AC	IoU
Parametric					
Pix2Pix [32]	1	✗	17.9	8.9	4.9
Non-Parametric					
Ours	1	✗	44.5	31.0	20.9
Ours	5	✓	58.2	41.2	31.4

Table 2. **Mask-RCNN Scores on COCO:** We use a pre-trained Mask-RCNN model [25] 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.

while our approach is most likable, there are still many situations where our approach produced undesirable outputs.

Cityscapes: Table 3 contrasts the performance of our approach with prior approaches [9, 32, 43, 54] that have specifically demonstrated on Cityscapes. Except Pix2Pix, we used publicly available results for this evaluation. Our approach is competitive to prior parametric and semi-parametric approaches with just 25 exemplar matches 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.

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

Method	#examples	Oracle	PC	AC	IoU
Parametric					
Pix2Pix [32]	1	✗	72.5	29.5	24.6
CRN [9]	1	✗	49.0	22.5	18.2
Pix2Pix-HD [54]	1	✗	79.0	43.3	37.8
Semi-Parametric					
SIMS [43]	1	✗	68.6	35.1	28.1
Non-Parametric					
Ours (top-25)	1	✗	67.1	38.0	30.5
Ours (top-25)	5	✓	71.3	39.6	32.4

Table 3. **PSP-Net Scores on Cityscapes:** We use a pre-trained PSP-Net model [59] to evaluate the quality of synthesized images (1024×2048). 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 exemplar matches to extract shapes and parts, our non-parametric approach is competitive to the parametric models and semi-parametric models.

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

[1] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *ECCV*, 2018. 1, 2

[2] Aayush Bansal, Yaser Sheikh, and Deva Ramanan. PixelNN: Example-based image synthesis. In *ICLR*, 2018. 2, 4

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 2009. 4

[4] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002. 3

[5] I. Biederman. *On the Semantics of a Glance at a Scene*. 3

[6] I. Biederman. Recognition by components: a theory of human image interpretation. *Pyschological review*, 94:115–147, 1987. 3, 5

[7] Oren Boiman and Michal Irani. Similarity by composition. In *NIPS*. 2006. 3

[8] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[9] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 1, 2, 6, 7, 8, 9

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 2, 3, 7

[11] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 6

[12] Santosh Kumar Divvala, Derek Hoiem, James H. Hays, Alexei A. Efros, and Martial Hebert. An empirical study of context in object detection. In *CVPR*, 2009. 3

[13] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *ICCV*, 2003. 1, 3

[14] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. 2001. 1, 3

[15] Alexei A. Efros and Thomas K. Leung. Texture synthesis by non-parametric sampling. In *ICCV*, 1999. 3, 5

[16] A. Faktor and M. Irani. Co-segmentation by composition. In *ICCV*, 2013. 3

[17] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 3

[18] William T. Freeman, Thouis R. Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Comput. Graph. Appl.*, 2002. 3

[19] Arnab Ghosh, Viveka Kulharia, Vinay P. Namboodiri, Philip H. S. Torr, and Puneet Kumar Dokania. Multi-agent diverse generative adversarial networks. In *CVPR*, 2018. 2, 7

[20] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. In *NIPS*, 2014. 1, 2

[21] Chunhui Gu, Joseph J Lim, Pablo Arbeláez, and Jitendra Malik. Recognition using regions. In *CVPR*, 2009. 3

[22] A. Gupta, A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, 2010. 3

[23] B. Hariharan, P. Arbelez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 3

[24] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Trans. Graph.*, 2007. 2, 3

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 7, 8

[26] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. Image analogies. *ACM Trans. Graph.*, 2001. 1, 2, 3

[27] 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 *NIPS*. 2017. 7, 8

[28] 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. 1

[29] D. Hoiem, A.A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006. 3

[30] Shiyu Huang and Deva Ramanan. Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters. In *CVPR*, 2017. 1

[31] Phillip Isola and Ce Liu. Scene collaging: Analysis and synthesis of natural images with semantic layers. In *ICCV*, 2013. 3

[32] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 2, 6, 7, 8, 9

[33] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 1

[34] Micah K. Johnson, Kevin Dale, Shai Avidan, Hanspeter Pfister, William T. Freeman, and Wojciech Matusik. CG2Real: Improving the realism of computer generated images using a large collection of photographs. In *IEEE Trans. Visualization and Computer Graphics*, 2011. 3

[35] Natasha Kholgade, Tomas Simon, Alexei Efros, and Yaser Sheikh. 3d object manipulation in a single photograph using stock 3d models. *ACM Trans. Graph.*, 2014. 4

[36] Jan J. Koenderink. *Solid Shape*. MIT Press, Cambridge, MA, USA, 1990. 3

[37] Jean-François Lalonde, Derek Hoiem, Alexei A. Efros, Carsten Rother, John Winn, and Antonio Criminisi. Photo clip art. *ACM Trans. Graph.*, 2007. 1, 2, 4

[38] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1

[39] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 2, 7

[40] David Marr. Vision: A computational investigation into the human representation and processing of visual information. 1982. 3

[41] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 3

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

[43] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *CVPR*, 2018. 3, 7, 8, 9

[44] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 2

[45] L. Roberts. Machine perception of 3-D solids. *PhD. Thesis*, 1965. 3

[46] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 2004. 4

[47] Bryan Russell, Alexei A Efros, Josef Sivic, William T Freeman, and Andrew Zisserman. Segmenting scenes by matching image composites. In *NIPS*, 2009. 1, 3

[48] Bryan Russell, William T Freeman, Alexei A Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 3

[49] Bryan Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 2008. 3

[50] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 3

[51] Brandon M. Smith, Li Zhang, Jonathan Brandt, Zhe L. Lin, and Jianchao Yang. Exemplar-based face parsing. In *CVPR*, 2013. 1, 2

[52] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 1

[53] Radim Tyleček and Radim Šára. Spatial pattern templates for recognition of objects with regular structure. In *GCPR*, 2013. 1, 2

[54] 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. 1, 2, 4, 7, 8, 9

[55] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 6

[56] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.* 3

[57] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 2

[58] 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 Trans. Graph.*, 2017. 4

[59] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 9

[60] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016. 4

[61] 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

[62] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2017. 2, 7