Disentangling Structure and Aesthetics for Style-aware Image Completion

Andrew Gilbert[1], John Collomosse[1 2], Hailin Jin[2], and Brian Price[2]

[1]Centre for Vision Speech and Signal Processing, University of Surrey
[2]Creative Intelligence Lab, Adobe Research

Abstract

Content-aware image completion or in-painting is a fundamental tool for the correction of defects or removal of objects in images. We propose a non-parametric in-painting algorithm that enforces both structural and aesthetic (style) consistency within the resulting image. Our contributions are two-fold: 1) we explicitly disentangle image structure and style during patch search and selection to ensure a visually consistent look and feel within the target image. 2) we perform adaptive stylization of patches to conform the aesthetics of selected patches to the target image, so harmonizing the integration of selected patches into the final composition. We show that explicit consideration of visual style during in-painting delivers excellent qualitative and quantitative results across the varied image styles and content, over the Places2 scene photographic dataset and a challenging new in-painting dataset of artwork derived from BAM!

1. Introduction

Image completion to repair defects or remove unwanted objects requires missing image data ('holes') to be filled in a visually plausible way. Most existing algorithms operate by copying and seamlessly blending patches from elsewhere in the image, hallucinating visually plausible texture to fill the hole [5, 4, 1]. This idea has been extended to sample patches from auxiliary image collections (AICs) [8] or domain-specific generative models [31, 10]. One advantage of using an AIC is greater flexibility and likelihood of finding a suitable match among the millions of images within the collection. Most methods focus on photographic images and accordingly are driven by structure or semantic similarity. However, this is not suitable for artistic images as these methods do not attempt to match on visual style.

In this paper we propose a novel AIC based image completion approach that *explicitly considers both structure and visual style* to deliver aesthetically superior in-painting results with a consistent visual look and feel, over a broad gamut of digital artwork. Specifically, we propose two technical contributions:

1) Style-aware Optimization. First, we explicitly factorize image appearance into structure and style through a deeply learned representation. This enables us to enforce style coherence within the patch search and selection optimization. Existing approaches focus upon only the structural plausibility of the completed image, considering this at both a low and high level. At a low level, patches are selected to minimize discontinuities in local edge and texture information in the image. At a high level, images are selected to ensure semantically similar patches (e.g., to in-fill a waterfall patches are sampled from images containing waterfalls). However, *neither of these structural constraints enforce a consistent visual aesthetic* between the patches selected and the image to be in-painted. For example, a visually plausible fill of a region of a watercolor painting would require the sampling of patches from a watercolor image. When drawing patches from AICs containing many millions of images [8], it is frequently possible to obtain patch candidates that match the object type and structure but have a very different aesthetic. Such visual inconsistencies are readily perceived by a human who expects a homogeneous aesthetic style within the completed image.

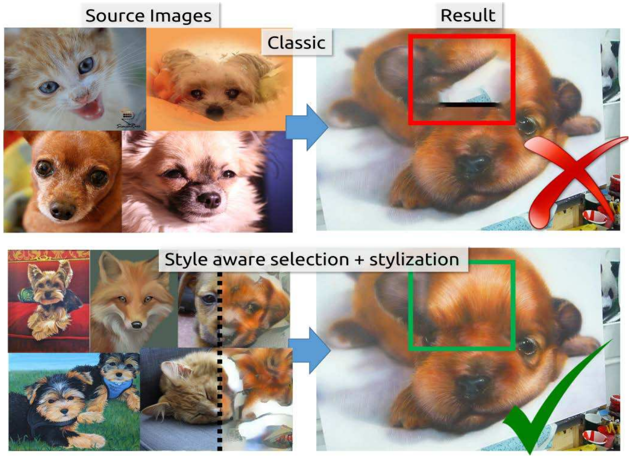

Figure 1. We propose enhancing image completion through explicit consideration of visual aesthetics (style) alongside structure and semantics. A deep convnet is used to disentangle patch structure and style, driving 1) style-aware patch search; 2) a style-aware optimization for patch selection; 3) stylization of patch content to enable seamless image completion with coherent visual aesthetic.

2) Adaptive Stylization. Second, our optimization adap-

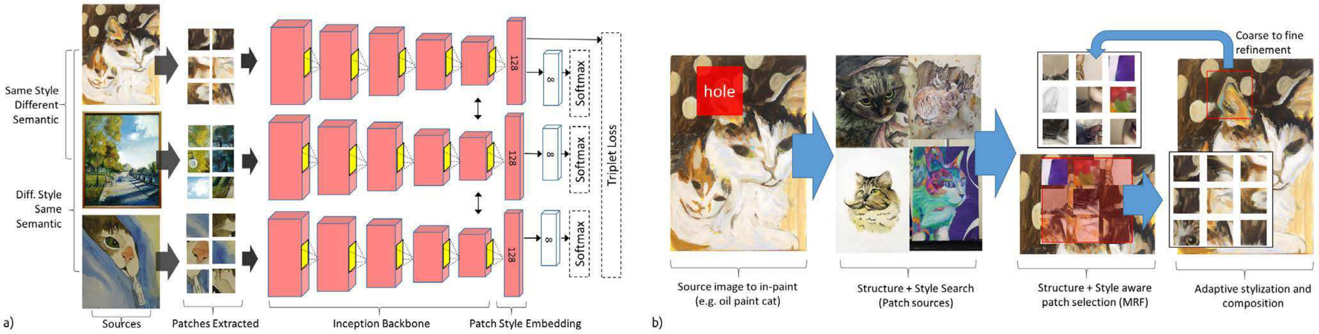

Figure 2. Overview of the proposed method. (a) Learning a local patch model to disentangle aesthetics and structure. (b) our proposed 3-stage algorithm for style aware image in-painting: 1) source image used as a query for a style-aware web-scale search for in-filling candidates; 2) patch aggregation and selection via MRF balancing structure and style; 3) adaptive stylization of patches using MRF weights.

tively stylizes patch content to conform their visual style to that of the destination image. Existing patch composition methods focus upon the removal of structural inconsistencies, e.g., using Poisson blending [21] or convolutional pyramids [6] to minimize discontinuities in the edges and texture. Although some approaches have explored basic color transfer [28] for patches, and generative texture (GAN) has been explored for narrow domains, no attempts have been made to harmonize the style of patches during their composition to create a homogeneous aesthetic in the filled region. This is important as patches may not exist within even large AICs that precisely match aesthetics of the image (or part thereof) being filled, especially when in-painting digital artwork. Thus stylization *broadens the gamut* of suitable content for in-painting; *bridging the gap between non-parametric and generative approaches* to image completion.

To incorporate an awareness of style (aesthetics) into both stages - patch selection and composition – our approach makes use of a feature embedding that can both quantify differences in style between patches (thereby influencing patch selection optimization) as well as drive stylization of chosen patches to finesse their appearance within the global composition. The abilities to disentangle structure and style similarity in both the search and selection of image patches, and to drive decision making around the formation of those patches (including their stylization) are the core novelties of this work. This enhances visual quality for automated image completion versus state of the art techniques, and diversifies image completion to the domain of digital artwork. We demonstrate the capability of our network in filling holes in stylized images using a novel dataset of artwork (BAM! [30]). We further show that our method excels in the photo in-painting task producing state-of-the-art results using the Places2 dataset [32]).

2. Related Work

Texture synthesis for image in-painting has been extensively studied. The earliest methods focused upon greedy per-pixel and patch-based iterative algorithms [21, 6, 29, 5] that sought to fill holes from their edges inward incrementally, sampling patches from elsewhere within the same image. The exploration of efficient representations for matching patches [14] and for efficient patch matching, e.g., using random propagation of a few good matches [1].

A common framework for texture synthesis is the Markov Random Field (MRF) in which the data term expresses the plausibility (via local image, e.g., the edge of texture properties) and the pair-wise term the spatial coherence. There have been several works exploring MRF formulations of this kind [16, 9, 19], including methods that source patches across multiple images (AICs) [8] and more recent methods that integrate deeply learned features (*e.g.* AlexNet/fc7) to constrain patch selection providing semantically coherent texture choice [24]. Although we also incorporate deeply learned features with an MRF formulation, our work uniquely considers visual style as a patch selection constraint to harmonize visual appearance.

Generative approaches *learn* the appearance of objects or scenes from large databases of images and leverage this to hallucinate missing regions. Pathak [20] presented a context encoder that understands semantics structure to complete the image holes. Yeh *et al*. [31] consider image completion as a constrained image generation problem, using a Generative Adversarial Network (GAN) model. Iizuka *et al*. [10] maintain both local and global consistency through a convolutional completion network and two context discriminator networks.

Our work straddles both the non-parametric sampling and generative paradigms. On the one hand, we adopt an AIC patch sampling approach driven by a deep model of structure and style, but on the other, we use the same embedding to modulate patch content beyond that available within the AIC using Neural Style Transfer (NST). Gatys [7] proposed the first NST work on the reconstruction of stylized images using a (generically trained) convnet using a loss function that offered independent control over semantic structure and visual style. Johnson *et al*. [12] proposed to pre-train a feed forward convnet over a broad set of images for each style to mitigate the computational burden of the NST. Structure and composition are typically preserved poorly within NST; an issue addressed by Li and Wand [17] who replaced the conventional Grammian approach to NST with a combination of a feed-forward CNN and a generative MRF (CNNMRF). Liao [18] and Upchurch [26] combined the concept of image analogy and deep features generated from a CNN to achieve

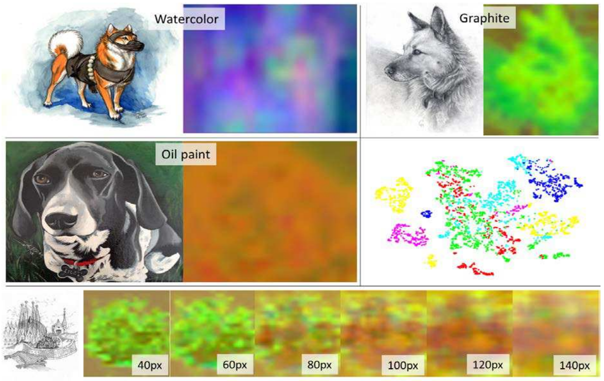

Figure 3. Dense visualizations (PCA) of patch style embedding for images of constant semantics but varying style (top) and varying spatial scales 40-140px (bottom). Inset: t-SNE embedding at 80px.

style transfer. In this work, we combine stylization and image completion using a single style embedding to exploit the style and structure of content,

3. Style-aware Image Completion

We propose a style-aware image completion algorithm comprising three stages (Fig. 2). Initially, a style- and structure-aware image search is performed to identify relevant images from a web-scale ($\sim 66.8M$ image) AIC from which raw patch data may be sampled (sec. 3.2.1). We then perform patch selection, the process of selecting from an over-complete set of patches available from this large set of auxiliary images (and the source image). Selection balances visual consistency, and semantic plausibility, expressed via a global objective function minimized through an MRF optimization performed at multiple scales (sec. 3.3). Patches are stylized during composition using the style distance term determined via the MRF patch selection, to minimize the difference between the style of the patch and source image, and so to remove artifacts and visual discontinuities (sec. 3.4). All stages of the pipeline are driven using a deeply learned feature embedding that disentangles structure and style.

3.1. Disentangling Patch Structure and Aesthetics

Explicit disentangling of structure and style improves the targeting of the image search to return content that more closely matches the aesthetics of the source image. It also enables reasoning about the style similarity of selected image patches, therefore determining the parameters of subsequent patch stylization. We learn a pair of functions $\{g_s(p), g_z(p)\}$, where p denotes an image patch, for feature embedding using two triplet convolutional neural networks (convnets): The first (g_s) learns aesthetic style similarity invariant to the semantics of image content, whilst the other (g_z) learns structural similarity invariant to style. Both networks are of homogeneous design, comprising a GoogLeNet (Inception-v3 [25]) backbone across all branches with a low-dimensional (128-D) bottleneck appended after the $pool5$ layer with all weights shared (Siamese). The feature embedding is available from this bottleneck layer. The network architecture is illustrated for the style embedding in Fig. 2a.

3.1.1 Style Embedding

Our coarse-to-fine in-painting process selects and stylizes patches at multiple resolutions (sec. 3.3-3.4) and therefore requires embeddings amenable to feature extraction at various scales. We therefore learn a set of style embeddings $\mathcal{S} = \{g_s^l\}$ for $l = [0, L]$ denoting half-octave intervals of square patch size from 40 to 160 pixels (px), with $i = 0$ indicating the full image (resized to 224px). Initially, we train each style embedding from scratch via discriminative (softmax) loss as a style classifier given a set of artistic images and photographs manually annotated into style categories, we use 88K images selected randomly from the BAM! (Behance Media) dataset [30], evenly partitioned into a set of 8 style categories comprising: watercolor; vectorart; 3D; graphite; pen; oil; comic; photo (see Fig. 3). The style network is then fine-tuned via hard negative mining using a triplet loss (eq. 1) by presenting image triplets in which the anchor a and positive branches p containing objects of the same style (*e.g.* watercolor) but differing content (e.g., a bike and a dog). The negative branch n comprises an image of differing style but an object similar to the anchor branch. The network minimizes the loss $\mathcal{L}$:

$$\mathcal{L}(a,p,n) = [m + |g_s^l(a) - g_s^l(p)|^2 - |g_s^l(a) - g_s^j(n)|^2]_+ \quad (1)$$

Where $m = 0.2$, a fixed constant related to the convergence and $[x]_+$ indicates positive values of x. A further $10K$ validation images were sampled from BAM! to evaluate the efficacy of the embedding via classification (mean average precision; mAP) score over scale. As expected, mAP falls with scale achieving 72.0, 65.6, 58.9 and 49.0 at 160px, 120px, 80px and 40px respectively. Despite this, performance is sufficient to extract style information at fine-resolution patch scales providing better definition (Fig. 3, lower) in turn improving patch selection during optimization.

3.1.2 Structure Embedding

We also learn the style-invariant structure embedding $g_z(.)$ via Inception-v3 branches pre-trained on ImageNet as a discriminative softmax problem followed by triplet refinement on artwork similar to $g_s(.)$. Precisely, in the case of $g_z(.)$ the roles of the positive and negative branches are reversed (*i.e.* anchor and positive branches receive similar semantics, but differing style, while the negative branch receives images matching the style but not the semantics of the anchor). Since BAM! in its current incarnation contains only nine semantic category annotations, to promote semantic generalization over artwork the network is subjected to a further fine-tuning stage over a broader dataset of 1M artworks from the website Behance (from which BAM! is derived). Fine-tuning is performed via an additional triplet refinement stage using string matches on keyword tags associated with artworks on the website to inform semantic relevance. Although in principal $g_z(.)$ could be trained over multiple scales, as per g_s^i we have found little performance benefit in doing so, perhaps due to the local nature of image

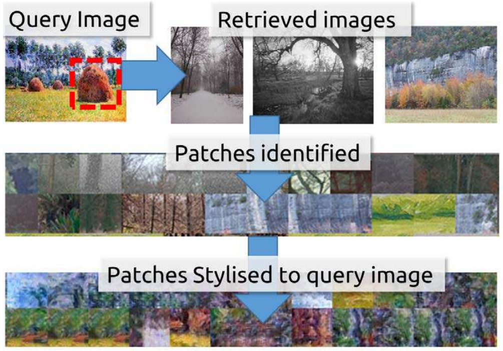

Figure 4. Top: Examples of a query (left) and retrieved images (right). Middle: Patches extracted densely. Bottom: Patch stylization performed to optimize aesthetic coherence when used to fill the hole. Sec. 4 discusses pre-stylization of patches vs. (optimally) adaptive stylization of patches via MRF optimization outcome.

structure, and thus a single embedding is learned for use in subsequent patch search and selection.

3.2. Patch Aggregation and Selection

Given a source image s containing a user-specified region ω to be in-painted, a set of candidate patches P need to be identified. Exhaustively exploring all patches within an AIC would prove intractable given the scale and diversity of collections available. In the spirit of [8], we automatically search via visual search a large dataset of images D to create a short-list of images from which to sample candidate patches.

3.2.1 Style and Structure aware Retrieval

To perform the search we index approximately $66.8M$ user-generated photos and artworks on Behance; a website for creative professionals. Each image $d \in D$ is forward passed through g_z and g_s^0 yielding a descriptor:

$$\mathcal{I}(d) = PQ\left(\left[\begin{array}{cc} g_z(d) & g_s^0(d) \end{array}\right], B\right) \tag{2}$$

where $PQ(.)$ indicates a product quantization (PQ) [11] of the 256-D search vector to a compact binary form (64 bits) using basis B learned over 0.5M images from D. The approach builds upon [2] in which style-aware search is performed using a convnet-learned projection of a similarly concatenated feature for sketch based retrieval. Here we employ PQ to scale over tens of millions of artworks, and return the top 200 based on a $k-$NN search on $||\mathcal{I}(s) - \mathcal{I}(d)||_2 \quad \forall d \in D$. The ability to search for images from which to pull texture, employing both structure (content) and style (aesthetic) constraints, is a unique and desirable property of our in-painting approach. Patches are sampled at multiple scales: 160×160px, 120×120px, 80×80px, and 40×40px (mapping to each iteration our multi-scale in-painting, sec. 3.3). The largest patches encode global structure, whilst smaller patches provide fine grain detail. We denote the unordered set of patches densely sampled from retrieved images as P.

3.3. Patch Selection over Learned Embeddings

Given the style and structure relevant patches, we propose a global optimization for filling the hole ω within s with patches maximizing visual plausibility and style coherence. We establish a regular, overlapping grid over ω each grid cell overlapping half of its neighbor. We consider hole filling as a labeling problem; a Markov Random Field (MRF) optimization is applied to select an optimal subset of the patches from candidate patch set P to label the grid. The optimization minimizes an energy function that balances the choice of patches to minimize deviation from three measures: content structure, style, and appearance (spatial coherence):

$$E(X) = \sum_{i\in\mathcal{V}} \psi_z(p_i) + \frac{1}{|N_i|} \sum_{i\in\mathcal{V}, j\in N_i} \psi_{ij}(p_i, p_j) + \sum_{i\in\mathcal{V}} \psi_s(p_i) \tag{3}$$

Where $\mathcal{V} = \{v_1, ..., v_n\}$ corresponds to the set of all grid cells, and $p_i \in P$ denotes the patch label associated with the i^{th} cell v_i, thus the energy term $E(X)$ evaluates putative mappings $X = \{v_i \mapsto p_i\} \quad \forall v_i \in \mathcal{V}$. N_i denotes the set of neighboring (4-connected) cells to v_i. The unary or data function $\psi_i(p_i)$ measures the deviation of the structure of patch p_i from the structured content in the source image (s) and is expressed via L_2 distance in the structure embedding (sec. 3.1.2) between the patch and image:

$$\psi_z(p_i) = ||g_z(p_i) - g_z(s)||_2 \tag{4}$$

The pairwise term $\psi_{ij}(p_i, p_j)$ measures spatial coherence of the patch neighborhood, through the sum of square difference (SSD) of pixel values in the overlap area between neighboring patches i, j (echoing the standard per pixel RGB difference used in the original GrabCut work [23] and derivatives). The tertiary term $\psi_s(p_i)$ encourages style coherence with local regions of the image. This is expressed as the L_2 distance within the style embedding (sec. 3.1.1):

$$\psi_s(p_i) = |g_s^l(p_i) - g_s^l(s)| + \frac{1}{|N_i|} \sum_{p_j \in \mathcal{N}_i} |g_s^l(p_i) - g_s^l(p_j)| \tag{5}$$

Thereby minimization of $E(X)$ encourages a spatially coherent in-painting both with respect to edge information (pairwise term) and local style coherence (tertiary term), while ensuring similar local semantic distribution (unary term). The impact of the structure (*e.g.* vs. SSD) and style terms in particular is studied in sec. 4. The proposed energy function takes a similar form to the Robust P^n model of [15] and a similar trick can be used to modify the energy term to take the form of a weighted average unary potential of patches. This definition is convenient as this spatially higher order term does not take multiple numbers of variables in the clique, and so can effectively be further merged to the unary term simplifying the energy function to a form solvable using standard alpha-beta expansion:

$$E(X) = \sum_{i,j\in\mathcal{V}} (\psi_z(p_i) + \psi_s(p_i, p_j)) + \sum_{i\in\mathcal{V}, j\in N_i} \psi_{ij}(p_i, p_j) \tag{6}$$

The MRF is solved iteratively at multiple scale levels $l = [0, L]$ using the corresponding set of style embeddings $g_s^l(.)$

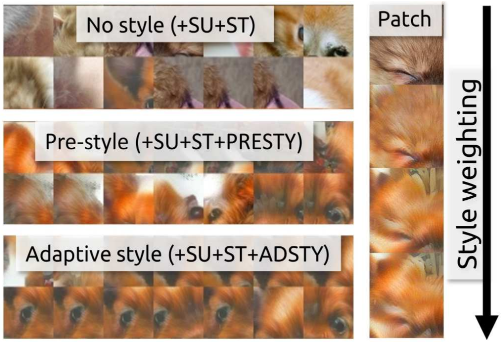

Figure 5. Patch Stylization. Left: Comparing patches without stylization (top), with pre-stylization (middle), and with adaptive stylization (bottom) for the in-painted dog in Fig 8. Right: Effect of increasing stylization weight on an adaptively stylized patch.

learned in Sec. 3.1.1. At each scale, s is updated and used as the input to eqs.4-5 such that finer scale iterations build upon the structure laid down by earlier, coarser iterations. At the initial (coarsest) iteration, $l = 0$ features derived from s in eqs.4-5 include the hole, encoded as zero (black) pixels.

3.4. Adaptive Patch Stylization

Our algorithm adaptively stylizes the set of selected patches X to harmonize patch content prior to compositing into s. Adapting neural style transfer (NST) [7] we extract a structure descriptor from the patch $\chi_z(p_i)$ and a style descriptor from the source image $\chi_s(s)$. We seek a modified patch p'_i such that $\chi_s(p') \cong \chi_s(s)$ and $\chi_z(p') \cong \chi_z(p_i)$. The degree to which style consistency between p'_i and s is enforced is driven by the tertiary term of the MRF:

$$\mathcal{L}_{sty}(p'_i) = |\chi_z(p'_i) - \chi_z(p_i)| - \alpha e^{-\psi_s(p_i)}|\chi_s(p'_i) - \chi_s(s)| \quad (7)$$

where $\alpha = 10^{-5}$, a scale normalization balancing structure and style terms. The solution for p'_i is initialized to p_i plus Gaussian noise, and loss term $\mathcal{L}_{\text{sty}}$ and minimized through back-propagation (ADAM) through a convnet. Following [7] $\chi_z(.)$ is obtained through forward-pass of a pre-trained (on ImageNet) VGG-19 sampling the $conv\_4$ layer and $\chi_s(.)$ through coarse style embedding $g^0_s(.)$. Thus the 'strength' of the stylization is governed by style similarity determined during MRF optimization. The stylized patches are composited into s via the gradient domain blending algorithm of Perez and Blake [21]. Fig 4 illustrates patches from retrieved images conformed to the style of s. In sec. 4.2 we contrast adaptive- vs. pre- stylization of retrieved images showing the latter to perform better. Adaptive stylization of patches is important for artwork where there is broad style diversity, and in-painted aesthetics must match.

4. Experiments and Discussion

We evaluate proposed algorithm over two datasets: 1) Places2 [32]; a dataset of photos commonly used for in-painting; 2) a new in-painting dataset sampled from a dataset of digital artwork 'Behance Artistic Media' (BAM!) [30]. Results are quantified via both a subjective user study and two objective metrics: structural similarity (SSIM) [27] and sliced Wasserstein distance (SWD) [13]. The latter is computed on patches sampled from the images; a Laplacian pyramid from a resolution of 16 × 16px is progressively doubled to 512 × 512px. 128 random descriptors formed of a 7×7px neighborhood are sampled from each level, and statistical similarity computed via sliced Wasserstein distance (SWD) — an approximation to Earthmover's distance [22]. Low SWD indicates the distribution of sampled patches between the in-painted image and ground truth are similar.

4.1. Comparative Evaluation

We compare against several contemporary baselines: PatchMatch [1] and Image Melding [3] which sample patches from within a single source image; an AIC approach 'Scene completion with millions of images' [8] (sec. 3.2.1); and a generative convnet approach 'Context encoder' [20]. Comparisons are made using published code.

A subset of 1000 digital artworks were sampled from BAM! across 8 artistic styles (sec. 3.1.1). Each image was re-sized to have longest side of 600px and a random mask defining the 'hole' for in-painting of side 150-250 pixels positioned at random. Representative results are shown in Fig 6, and Fig 7 compares the proposed method to each baseline, and results are quantified in Tbl. 1 (upper). Existing baselines suffer from blurring and failure to in-paint details such as the dog eye or cat ear. Our method reduces false positives that may match structure but not visual style of the source. Adaptive stylization harmonizes each patch's appearance, further reducing artifacts.

We compare against the same baselines using scenic photographs in Places2, plus a recent GAN in-painting work [10] (ImgComp.) for which no code or quantified results are released. Fig 9 presents a visual comparison, and quantitative results are given in Tbl.1 over the image set included in [10] with the same mask regions published in that work. On this photo dataset we exceed the state of the art, including a GAN [10] requiring months of GPU training (but realtime inferernce). The other methods struggle with the difficulty of these images creating blurred regions (ImgMelding, Context Encoder) that detract from the visual quality and coherence of the image. Whilst PatchMatch produces reasonable visual results, they are structurally incorrect in these examples.

Run-time varied across baselines for a 600px image: a few seconds (PatchMatch, ContextEncoder) to a few minutes (Millions of Images) to 1.5 hours (Image Melding). Solution of our MRF takes less than 30 seconds, with adaptive NST taking a few minutes and representing the bottleneck.

4.2. Ablation Study

Our ablation study cumulatively enables each of our individual contributions on top of a classic baseline for in-painting. BAM results are presented in Tbl. 1 (lower) and visualized in Fig. 8, and the results are presented in Tbl 3 for the Places2 examples from [10]. We first consider our

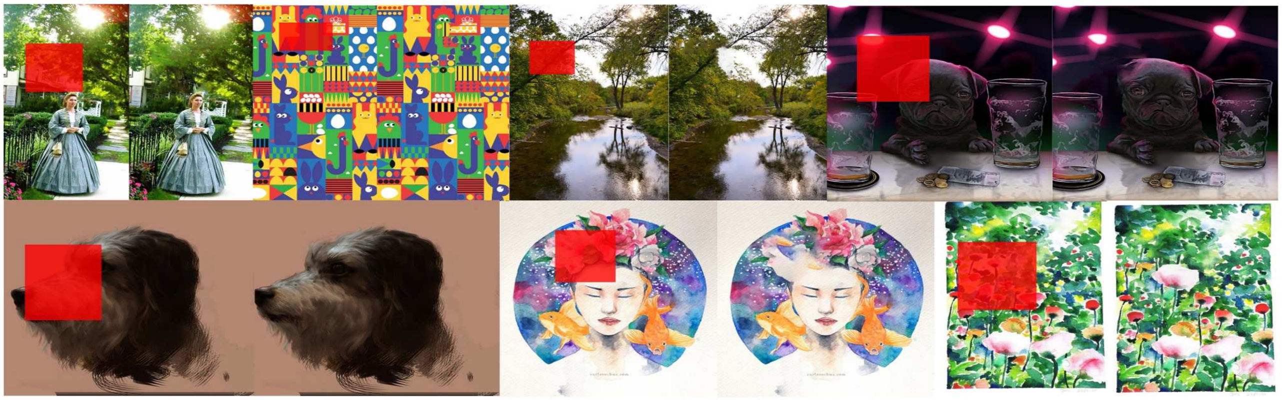

Figure 6. Representative results of the proposed style- and structure- aware image completion algorithm over photos and digital artwork (Places2, BAM). Source (left) and result (right); in-painted region ('hole') highlighted in source.

Method	Style																	
	3D		Comic		Graphite		Oil		Photo		Pen Ink		Vector		WaterColor		Mean	
	SSIM	SWD	SSIM	SWD	SSIM	SWD	SSIM	SWD	SSIM	SWD	SSIM	SWD	SSIM	SWD	SSIM	SWD	SSIM	SWD
Million Image [8]	0.85	2.34	0.87	2.41	0.89	2.30	0.84	2.37	0.86	2.41	0.84	2.30	0.9	2.31	0.84	2.35	0.86	2.35
PatchMatch [1]	0.86	2.33	0.91	2.20	0.91	2.19	**0.91**	**2.14**	**0.91**	**2.30**	0.88	2.23	**0.94**	**2.16**	0.91	2.26	0.91	2.23
PatchMatch[1]+NoStyle	0.87	2.32	0.91	2.20	0.91	2.19	**0.91**	**2.14**	**0.91**	**2.30**	0.88	2.23	**0.94**	2.16	0.91	2.26	0.91	2.22
PatchMatch[1]+PRESTY	0.88	2.31	0.91	2.21	0.91	2.19	**0.91**	**2.13**	**0.91**	**2.30**	0.90	2.21	**0.94**	2.16	0.91	2.26	0.91	2.22
ImgMelding [3]	0.81	2.48	0.88	2.41	0.86	2.28	0.87	2.29	0.84	2.39	0.85	2.30	0.89	2.32	0.83	2.37	0.85	2.36
ImgMelding[3]+NoStyle	0.81	2.48	0.88	2.41	0.86	2.28	0.87	2.28	0.84	2.39	0.85	2.31	0.89	2.32	0.83	2.37	0.85	2.36
Context Encoder [20]	0.86	2.27	0.82	2.26	0.91	2.29	0.83	2.24	**0.91**	2.30	0.81	2.31	0.9	2.31	0.84	2.36	0.86	2.29
Baseline (NoStyle)	0.85	2.39	0.88	2.27	0.89	2.40	0.84	2.41	0.85	2.35	0.85	2.28	0.93	2.28	0.89	2.38	0.87	2.35
+SU	0.86	2.35	0.89	2.23	0.89	2.35	0.84	2.41	0.86	2.34	0.85	2.28	**0.94**	2.18	0.89	2.38	0.88	2.32
+SU+ST	0.87	2.34	0.89	2.23	0.91	2.27	0.85	2.39	0.86	2.34	0.85	2.28	**0.94**	2.18	0.89	2.37	0.88	2.30
+SU+ST+PRESTY	0.91	2.33	**0.92**	**2.21**	0.90	2.19	0.89	2.19	0.90	**2.30**	0.88	2.27	**0.94**	**2.17**	0.93	2.26	0.91	2.24
+SU+ST+ADSTY (Ours)	**0.94**	**2.17**	0.91	**2.21**	**0.92**	**2.17**	**0.91**	2.15	**0.91**	**2.30**	**0.93**	**2.14**	**0.94**	**2.17**	**0.94**	**2.25**	**0.93**	**2.19**

Table 1. Structural image similarity (SSIM) vs. the ground truth for BAM. SSIM, higher is better, SWD ($\times 10^2$) lower is better

MRF patch selection without style – eq. 3 with unary and pairwise terms only – using SSD for $\psi_z(.)$ rather than $g_z(.)$ as structure embedding; result *no patch style (baseline)*. The result *structure unary* (+SU) uses $\psi_z(.)$ whilst *style term* (+ST) enables the tertiary term $\psi_s(.)$ for style. We also explore pre-stylization of P using NST [7] prior to solving the MRF *pre styled images* (+SU+ST+PRESTY). We contrast this with our full proposed approach which *adaptively stylizes* (+SU+ST+ADSTY) patches based on $\psi_s(.)$.

Our method outperforms all baselines for the Places2 dataset, and for the majority of BAM data with PatchMatch and Context encoder performing similar to the proposed method for photo content in BAM. Patches used in the +SU+ST variant have similar appearance but may mismatch structure or visual style of s, while the +SU+ST+PRESTY patches some are similar in aesthetic but not self-consistent within the image. The difference between the two is further illustrated in the in-painted dog image (Fig. 8) and visualized in Fig 5. Further inefficiency exists with +SU+ST+PRESTY is that the stylization occurs before the MRF optimization has occurred, meaning all possible patches have to be stylized. Stylization during compositing allows a margin for error since patches that deviate slightly from the aesthetics of their neighbors can be finessed into a homogeneous style.

4.3. Perceptual User Study

We conducted a study via Amazon Mechanical Turk (AMT) to compare performance of our method versus the 3 most promising baselines, and 2 most promising ablations. 300 images sampled from BAM were manipulated (as sec. 4.1) to remove a random region of interest. We concatenated results from the 6 methods in random order and presented them to 30 participants, gathering in total 9k annotations from 243 unique users. Participants were asked to 'identify the highest quality image' without sight of the ground truth. A consensus threshold of $\frac{1}{3}$ of the total votes for each image was used to disregard results that failed to reach consensus. Tbl. 2 reports the preferences expressed. Our approach significantly outperforms the ablated variants and existing baselines with results echoing Tbl.1 trends.

Method	Style								
	3D	Comic	Grphte	Oil	Photo	Pen Ink	Vect	WtrClr	Tot
IM[3]	0.00	0.89	1.04	0.00	0.00	0.76	0.00	0.89	3.67
PM [1]	1.03	1.39	1.03	0.00	**4.87**	0.00	1.67	2.35	12.45
CE [20]	0.00	1.43	2.23	0.00	4.03	0.00	0.00	0.00	7.66
+SU	0.00	1.02	0.78	2.32	0.00	1.89	0.00	0.00	6.01
+PRESTY	4.43	0.79	0.00	1.89	0.00	0.00	0.00	6.34	12.04
Ours	**10.4**	**5.43**	**7.27**	**6.37**	3.85	**7.10**	**8.34**	**10.92**	**59.34**

Table 2. Perceptual user study on BAM: preference across 6 methods (as %, 3.s.f.). Users asked for the "highest quality" image.

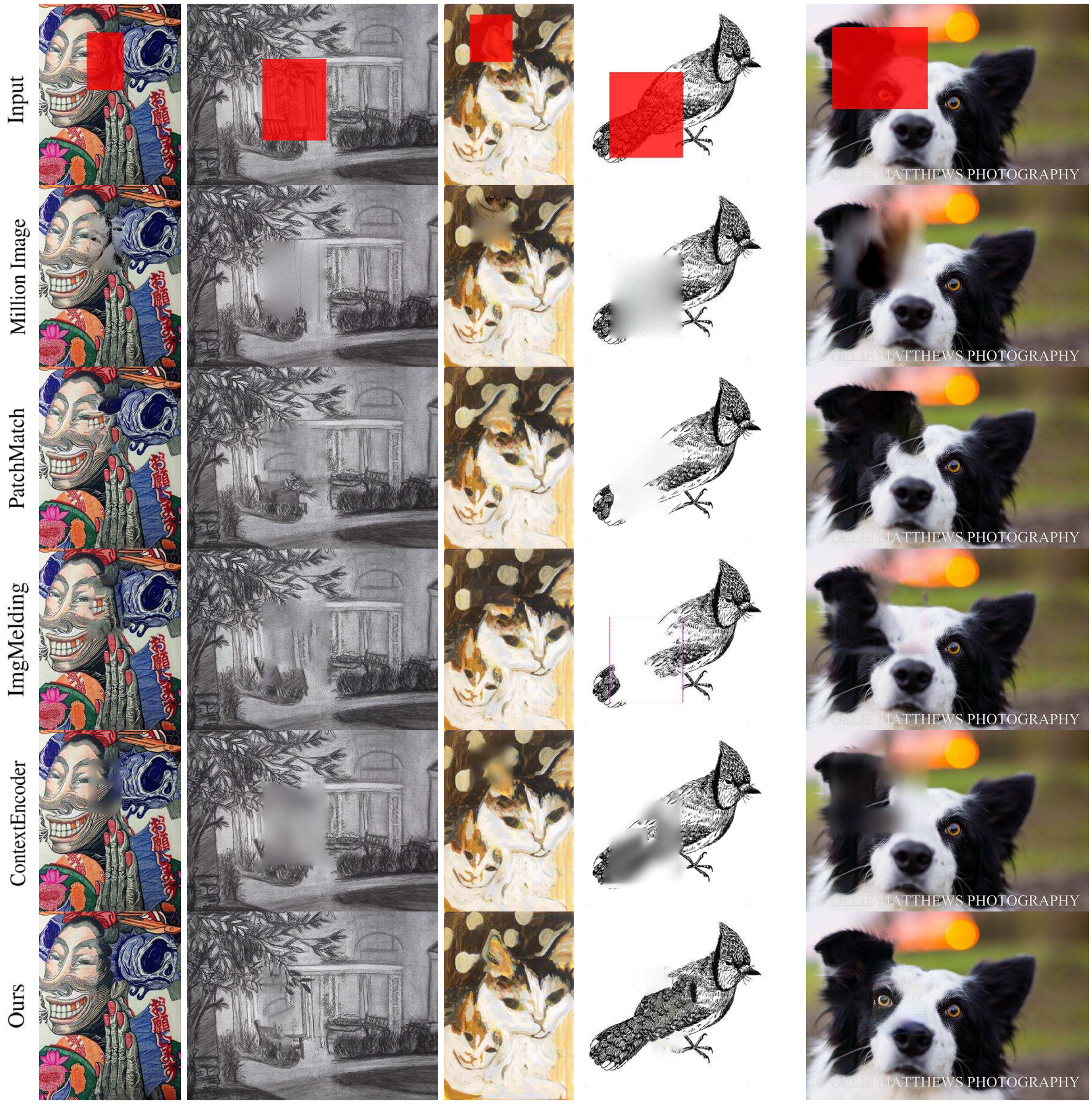

Figure 7. Comparison with existing works, AIC million image [8], PatchMatch [1], Img Melding [3], and Context Encoder [20]. Existing works produce reasonable results but with excessive blurring or missing key details absent in the image due to inability to identify patches that are coherent in structure and style. Our approach identifies patches from similar structure and style images to hallucinate detail that is not present in the image, and also then restyle the patches further to ensure a visually coherent result. More results in suppl. material

4.3.1 Failure cases

Fig 10 highlights some failure cases encountered by the proposed method. Adopting NST [7] for stylization limits performance on styles that are not well transferred by that method. In Fig. 10(a) the face isn't fully completed, and in general fine media types such as pen-ink artworks are harder for NST, and so our pipeline, to handle. Similarly the short-listing of retrieved images via visual search presents a limitation when quality of retrieval is lower as we retrieve without stylisation. Fig 10(b) shows a car failing to in-paint as the search focused on 3D buildings rather than cars. Fig 10(c) fails due to the sofa style and structure being retrieved but with insufficient detail to complete the stylized people in the scene.

5. Conclusion

We have presented a novel algorithm for style-aware image completion. The core novelties of our method are a deep embedding for visual structure and style that enables: 1) the consideration of visual style in the search and selection of

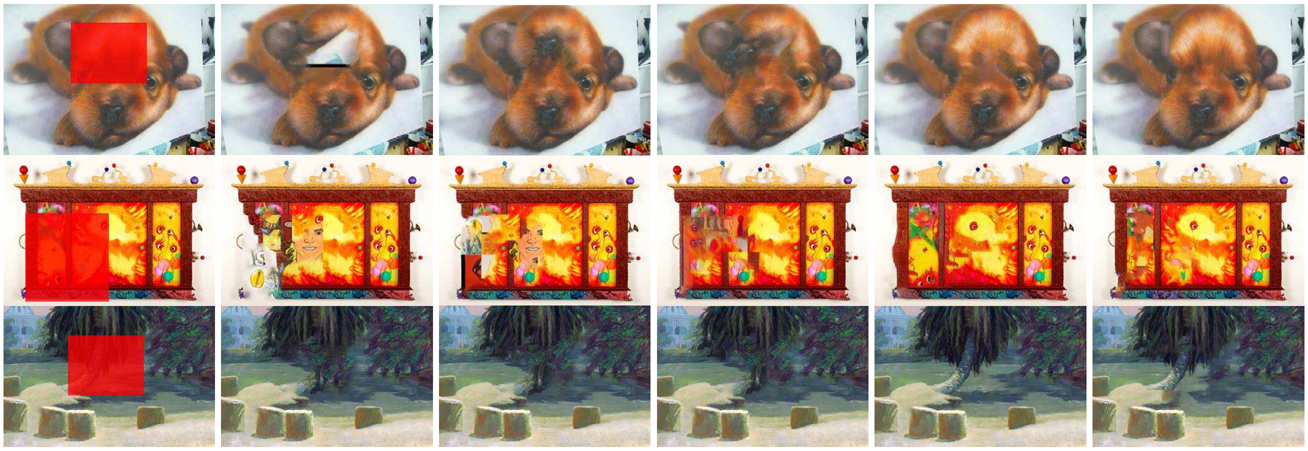

Figure 8. Ablation study cumulatively enabling individual style aware components of the approach to demonstrate their impact.

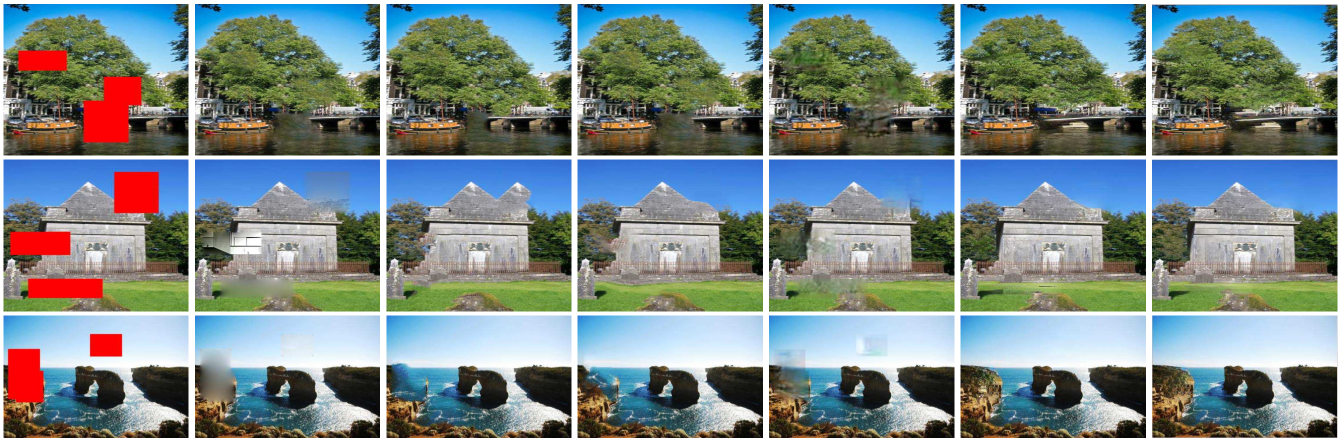

Figure 9. Qualitative comparison of five baselines vs. the proposed approach on the Places2 dataset.

	Approach							
	IM[3]	PM [1]	CE [20]	GL [10]	+SU	+ST	+PRESTY	+ADSTY
SSIM	0.76	0.88	0.74	0.89	0.79	0.81	0.85	**0.90**
SWD	2.62	2.60	2.50	2.47	2.60	2.51	2.55	**2.47**

Table 3. Method and baselines over Places2 results in [10]. SSIM, higher is better, SWD ($\times 10^2$) lower is better. Ours is *+ADSTY*

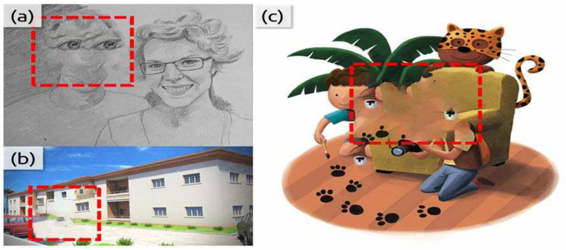

Figure 10. Illustrative failure cases, due to limitations in content availability or stylization (c.f. Sec. 4.3.1).

patches; and 2) the neural stylization of patch content to promote consistency of visual style within the in-painted image. As such our approach bridges the gap between patch-based and generative approaches to image in-painting. Our algorithm delivers results quantitatively and qualitatively superior to the state of the art on photos in Places2 [32], and non-photo artwork in a subset of BAM! [30]. We have demonstrated the independent value of our style-aware optimization and stylization contributions through ablation studies. Future work might include pursuing a fully generative approach to in-painting, *e.g.* similar to [10] but enabling style-aware patch hallucination and so generality beyond photographic data for GAN-based approaches.

Acknowledgements

This work was supported by a charitable donation from Adobe Systems, and also in part by the InnovateUK TotalCapture project grant agreement 102685 and the Visual Media project (EU H2020 grant 687800).

References

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing,. In *Proc. ACM SIGGRAPH*, 2009. 1, 2, 5, 6, 7, 8

[2] J. Collomosse, T. Bui, M. Wilber, C. Fang, and H. Jin. Sketching with style: Visual search with sketches and aesthetic context. In *CVPR'17*, pages 488–496, 2017. 4

[3] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen. Image melding: combining inconsistent images using patch-based synthesis. In *ACM TOG*, 2012. 5, 6, 7, 8

[4] A. Efros and W. Freeman. Image quilting for texture synthesis and transfer. In *Proc. SIGGRAPH*, 2001. 1

[5] A. Efros and T. Leung. Texture Synthesis by non-parametric sampling. In *Proc. Intl. Conf. on Computer Vision (ICCV)*, 1999. 1, 2

[6] Z. Farbman, R. Fattal, and D. Lischinski. Convolution pyramids. *ACM SIGGRAPH Asia*, 2011. 2

[7] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 2, 5, 6, 7

[8] J. Hays and A. A. Efros. Scene completion using millions of photographs. In *ACM Transactions on Graphics (TOG)*. ACM, 2007. 1, 2, 4, 5, 6, 7, 8

[9] K. He and J. Sun. Statistics of patch offsets for image completion. In *Euro. Conf. on Comp. Vision (ECCV)*, 2012. 2

[10] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and Locally Consistent Image Completion. *ACM Transactions on Graphics (Proc. of SIGGRAPH 2017)*, 36(4):107:1–107:14, 2017. 1, 2, 5, 8

[11] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2011. 4

[12] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 2

[13] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. 5

[14] F. Klose, O. Wang, J.-C. Bazin, M. Magnor, and A. Sorkine-Hornung. Sampling based scene-space video processing. In *ACM Transactions on Graphics*, 2015. 2

[15] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009. 4

[16] V. Kwatra, A. Schodl, I. Essa, G. Turk, and A. Bobick. Graphcut textures: Image and video synthesis using graph cuts. *ACM Transactions on Graphics*, 3(22):277–286, 2003. 2

[17] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2479–2486, 2016. 2

[18] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017. 2

[19] Y. Liu and V. Caselles. Exemplar-based image inpainting using multiscale graph cuts. *IEEE Trans. on Image Processing*, pages 1699–1711, 2013. 2

[20] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *CVPR'16*, 2016. 2, 5, 6, 7, 8

[21] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. In *ACM Transactions on graphics (TOG)*. ACM, 2003. 2, 5

[22] J. Rabin, J. Delon, and Y. Gousseau. Circular earth movers distance for the comparison of local features. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008. 5

[23] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*. ACM, 2004. 4

[24] O. Sendik and D. Cohen-Or. Deep correlations for texture synthesis. *ACM Transactions on Graphics (TOG)*, 36(5):161, 2017. 2

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR'15*, pages 1–9, 2015. 3

[26] P. Upchurch, N. Snavely, and K. Bala. From a to z: supervised transfer of style and content using deep neural network generators. *arXiv preprint arXiv:1603.02003*, 2016. 2

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[28] L. Wei and M. Levoy. Fast texture synthesis using tree-structured vector quantization. In *Proc. SIGGRAPH*, 2000. 2

[29] L.-Y. Wei and M. Levoy. Fast texture synthesis using tree-structured vector quantization. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 479–488. ACM Press/Addison-Wesley Publishing Co., 2000. 2

[30] M. J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, and S. Belongie. Bam! the behance artistic media dataset for recognition beyond photography. *arXiv preprint arXiv:1704.08614*, 2017. 2, 3, 5, 8

[31] R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2016. 1, 2

[32] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016. 2, 5, 8