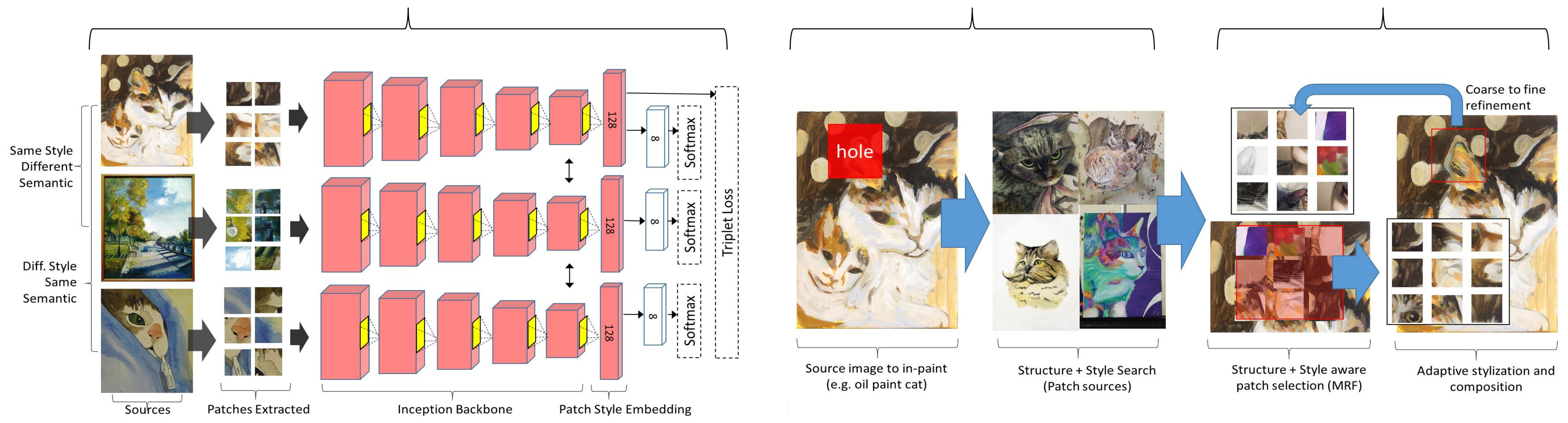
Disentangling Structure and Aesthetics for Style-aware Image Completion

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

[1] Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, UK.
[2] Adobe Research, Creative Intelligence Lab (CIL).

Introduction - Image Completion

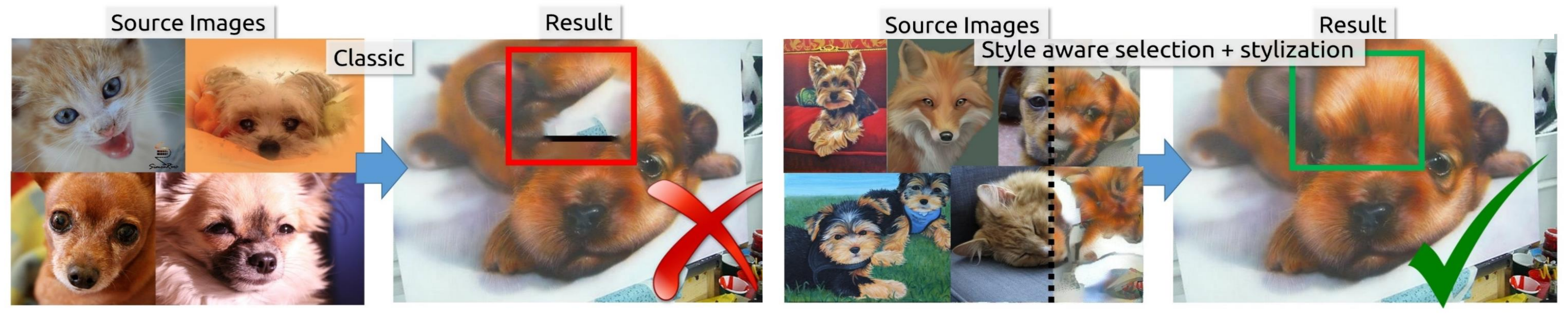
Image completion (or "in-painting") enables the removal of unwanted objects of artefacts in images. Most prior work operates by copying patches from elsewhere in the same image, or from auxiliary image collections (AICs), so as to hallucinate plausible texture to in-paint the unwanted regions. Previous work focused on the in-painting of photographic images only, with patch selection and coping driven by structure or semantic similarity. Our novel contribution is a novel AIC based image completion approach that explicitly enforces both structural and style (aesthetic) consistency in the patch selection process, and adaptively stylizes patches for aesthetic consistency during the copying process.



Disentangling Patch Structure and Style

A triplet convnet is trained to disentangle patch structure and style, driving:

1. **Style and structure aware visual search** for candidate patches in the AIC;
2. A **style-aware global optimization** for patch selection;
3. **Adaptive stylisation** of patch content to enable seamless image completion

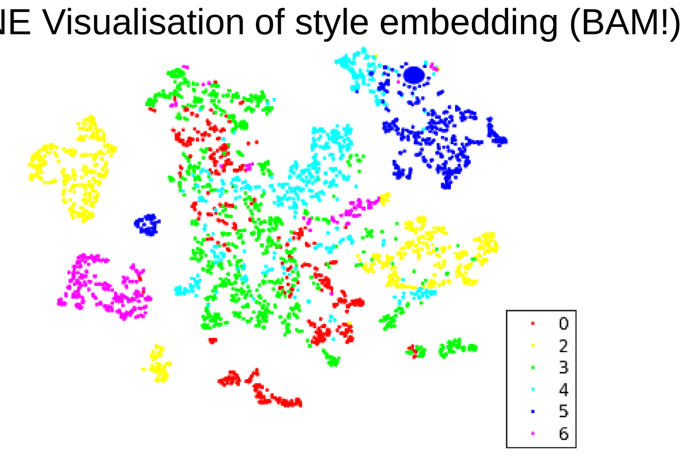


Candidate Patch Search

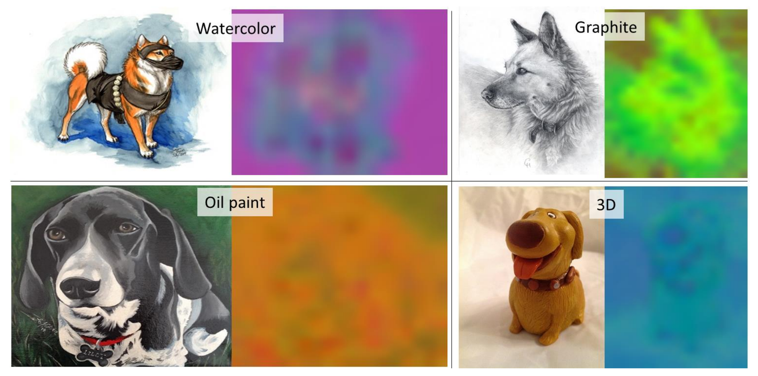
A style and structure aware image search is performed to identify relevant images from ~ 66.8M images from which raw patch data may be sampled

Learnt feature embedding using two triplet Inception-v3 convnets:

T-SNE Visualisation of style embedding (BAM!)



Style embedding as heatmap over images



Returns the top 200 images from 68M images on Behance; constrained by both structure (content) and style constraints,

Style-aware optimization for patch selection

Given the style and structure relevant patches, we propose a global optimization for filling the hole with patches maximizing visual plausibility and style coherence

E(X) = **Patch Structure** (data term) + **Spatial Coherence** (pairwise term) + **Style Coherence** (data term)

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

Patch Structure: measures the deviation of the structure of patch p_i from the structured content in the source image, s

$$\psi_z(p_i) = ||g_z(p_i) - g_z(s)||_2$$

Spatial Coherence: The pairwise term $\psi_{ij}(p_i p_j)$ measures spatial coherence of the patch neighbourhood, through the sum of square difference (SSD) of pixel values in the overlap area between neighbouring patches i, j

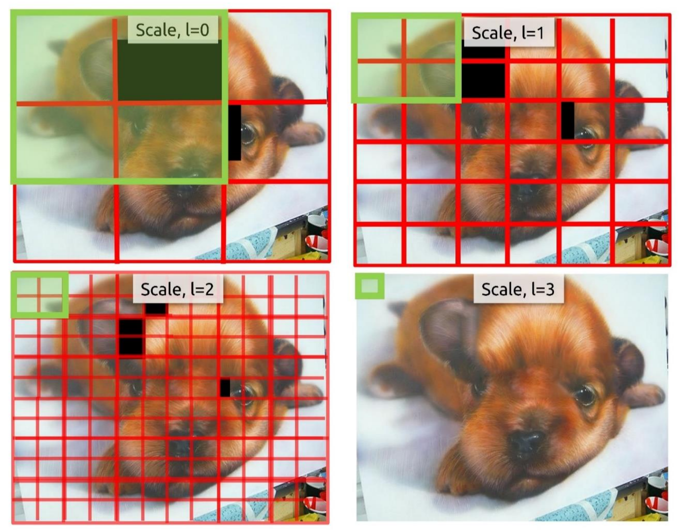
Style Coherence: Encourages style coherence in local regions of the image. This is expressed as the L2 distance within the style embedding

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

Coarse to fine

The MRF is solved iteratively at multiple scales.

Grid of overlapping patches

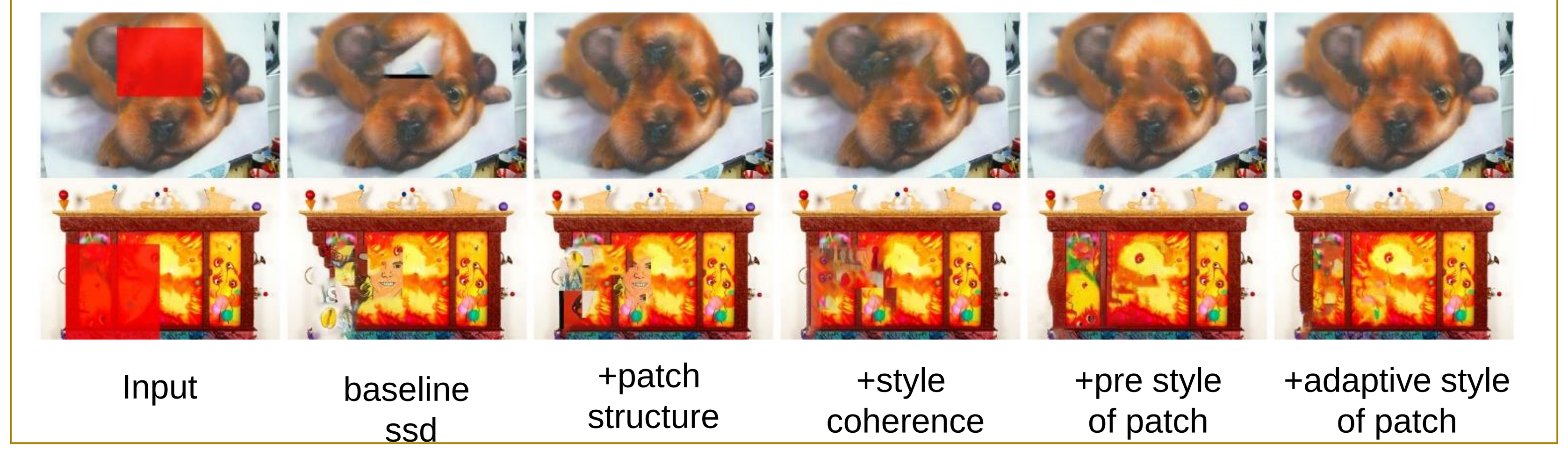


Adaptive Stylization

Adaptively stylizes the set of selected patches X to harmonize patch content prior to compositing into s

Ablation study

Cumulatively enables each of our individual contributions on top of a classic baseline for inpainting



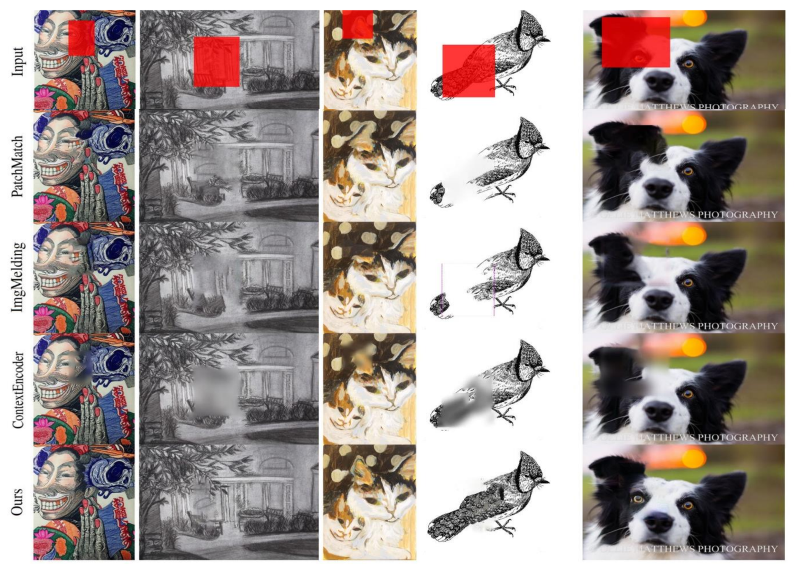
Evaluation

We evaluated the approach using two large image datasets:

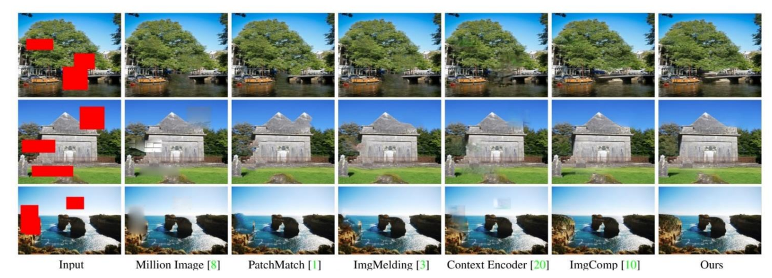
1) **Places2** a dataset of photos commonly used for image completion;
2) **Behance Artistic Media (BAM!)** a new in-painting dataset sampled from a website of publicly shared artwork from creative professionals

We compare against several contemporary baselines: PatchMatch [1], Image Melding [3], Efros et al./Million Image AIC [8] and Context Encoder [20].

Behance (BAM!)



Places2 Benchmark



Method	3D SSIM	3D SWD	Comic SSIM	Comic SWD	Graphite SSIM	Graphite SWD	Oil SSIM	Oil SWD	Photo SSIM	Photo SWD	Pen Ink SSIM	Pen Ink SWD	Vector SSIM	Vector SWD	WaterColor SSIM	WaterColor SWD	Mean SSIM	Mean 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

Additional Inpainting Results

Illustrating patch selection and stylization