Self-supervised disentangled representation learning of artistic style through Neural Style Transfer

[1]Dan Ruta, [1]Gemma Canet Tarrés, [1]Alexander Black, [1]Andrew Gilbert, [1,2]John Collomosse

[1]CVSSP University of Surrey, [2]Adobe Research

Background:

We set out to create a style representation learning model with far less entanglement of content information.

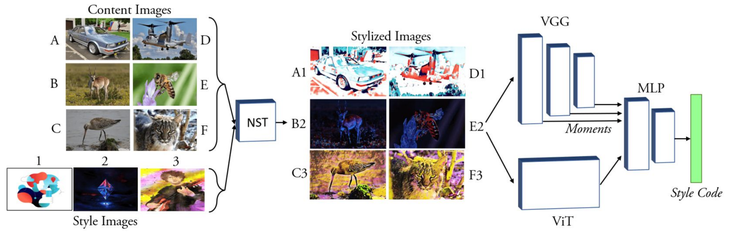
We use Neural Style Transfer (NST) to measure and drive the learning signal and achieve state-of-the-art representation learning on explicitly disentangled metrics. We show that strongly addressing the disentanglement of style and content leads to large gains in style-specific metrics, encoding far less semantic information and achieving state-of-the-art accuracy in downstream style matching (retrieval) and zero-shot style tagging tasks

BAM-FG

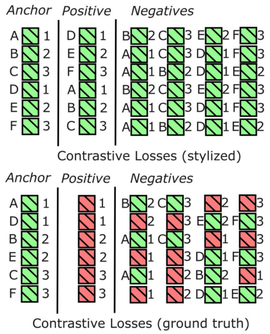
Synthetic

Contribution #1 - Novel training methodology for representation learning of style:

- We use NST dynamically during the training process to generate style-consisted data with randomized content subject matter
- We create invariance to the content, while training for style consistency
- We use two branches in our model:
  - A descriptive ViT branch with global attention
  - A VGG branch over which we can compute moment statistics, for all layers, similar to the original ALADIN design



- We use contrastive learning over positive data pairs of similar styles within a batch, amongst stylized results
- The negative samples are the other remaining samples stylized in that batch
- We also use the original style images as positives



Contribution #2 - SOTA style representation learning with enforced disentanglement:

- We achieve state-of-the-art style entanglement, as measured through image retrieval metrics
- Our method does not encode content information as well, also as measured through image retrieval, but over content images

Model	NST learning signal			NeAT test set		PAMA test set		SANet test set		*Average values*		CAST test set	
	NeAT	PAMA	SANet	mAP	IR-1	mAP	IR-1	mAP	IR-1	mAP	IR-1	mAP	IR-1
ALADIN-ViT	-	-	-	16.823	0.270	9.9964	0.0575	12.493	0.0599	13.104	0.129	39.541	0.960
Ours (ViT)	✓			85.306	66.308	51.012	15.303	67.525	28.423	67.948	36.678	39.472	**10.765**
Ours (ViT)		✓		69.226	23.415	62.886	20.628	51.934	6.393	61.349	16.812	24.563	3.608
Ours (ViT)			✓	80.230	46.466	56.215	18.738	74.621	35.693	70.355	33.632	38.171	8.895
Ours (ViT)	✓	✓		84.997	59.650	68.468	30.563	67.021	23.443	73.495	37.885	38.864	8.995
Ours (ViT)	✓		✓	85.052	64.993	53.657	17.688	77.410	46.408	72.040	**43.030**	**39.880**	10.715
Ours (ViT)		✓	✓	77.056	36.413	64.596	23.048	67.229	22.835	69.627	27.432	32.523	5.243
Ours (ViT)	✓	✓	✓	83.915	58.900	67.484	29.745	74.755	34.460	**75.385**	41.035	39.688	9.010

Tested through style image retrieval on synthetic test created with several methods. We ablate the NST methods used during training. We include a hold-out NST method as test set (CAST)

Model	Dataset	NeAT test set		PAMA test set		SANet test set		*Average values*		CAST test set	
		mAP	IR-1	mAP	IR-1	mAP	IR-1	mAP	IR-1	mAP	IR-1
ALADIN	BAM-FG	59.549	8.085	38.423	1.7025	48.712	3.560	48.895	4.449	14.066	0.145
⟶ Fused	BAM-FG	53.941	4.485	32.686	0.550	42.592	2.395	43.073	2.477	12.716	0.103
ALADIN-ViT	BAM-FG	16.823	0.270	9.996	0.058	12.493	0.060	13.104	0.129	39.541	0.960
SAE	BAM-FG	51.600	16.100	28.500	4.000	28.814	4.643	36.305	8.248	24.001	3.622
Ours	BAM-FG	85.955	58.108	67.699	24.967	74.355	27.154	76.003	36.743	45.352	8.963
Ours	BBST-4M	90.965	69.523	80.861	42.803	84.953	45.258	**85.593**	**52.528**	49.336	9.003
Ours + SS	BBST-4M	90.224	69.950	79.053	39.638	82.308	38.403	83.862	49.330	**60.811**	**14.640**
Ours + SS, no NST	BBST-4M	11.801	0.665	5.080	0.070	8.554	0.475	8.478	0.403	4.097	0.098

Style image retrieval compared to baselines - higher value is better

Model	Dataset	NeAT test set		PAMA test set		SANet test set		*Average values*		CAST test set	
		mAP	IR-1	mAP	IR-1	mAP	IR-1	mAP	IR-1	mAP	IR-1
ALADIN	BAM-FG	5.547	0	8.523	0	4.642	0	6.237	**0**	23.279	0.045
⟶ Fused	BAM-FG	11.008	0	19.239	0.013	8.920	0	13.056	0.004	41.015	0.183
ALADIN-ViT	BAM-FG	15.058	0.023	10.081	0.028	8.097	0.003	11.079	0.018	7.485	**0.010**
SAE	BAM-FG	2.198	0.003	3.815	0.005	2.758	0	2.924	0.003	5.354	**0.010**
Ours	BAM-FG	1.523	0	1.575	0	1.630	0	1.576	**0**	6.974	0.063
Ours	BBST-4M	1.491	0	1.427	0	1.652	0	1.523	**0**	7.463	0.070
Ours + SS	BBST-4M	1.381	0	1.461	0	1.629	0	**1.490**	**0**	**4.626**	0.018
Ours + SS, no NST	BBST-4M	10.789	0.355	12.034	0.345	5.969	0.010	9.597	0.237	31.687	3.443

Content image retrieval compared to baselines - lower value is better, to show the lack of learning content image features.

We also test with non-synthetic data, by first cropping high resolution data into smaller crops. This cropping breaks apart semantic content, slightly avoiding entanglement in the data

We still compare favourably, even when evaluating with non-synthetic data. The gaps are smaller, as there is still some content/style entanglement, despite the cropping.

Model	Dataset	Real image evaluation	
		mAP	IR-1
ALADIN	BAM-FG	0.519	0.840
⟶ Fused	BAM-FG	0.539	0.380
ALADIN-ViT	BAM-FG	0.483	9.450
SAE	BAM-FG	0.503	1.422
Ours	BAM-FG	0.510	8.040
Ours	BBST-4M	0.566	13.470
Ours + SS	BBST-4M	**0.575**	**14.060**
Ours + SS, no NST	BBST-4M	0.463	6.63

Contribution #3 - SOTA multi-modal vision/language representation learning of artistic style

- We can use our model in downstream tasks such as multimodal vision/language representation learning - again in the domain of artistic style
- By itself, our model does not achieve SOTA on the StyleBabel test set - we hypothesize this is due to entanglement in the StyleBabel data itself
- We do achieve SOTA when fusing our model with ALADIN-ViT, the previous SOTA method on StyleBabel, inspired by experiments in the paper of the original ALADIN model

ALADIN-ViT	neon-like, warm toned lighting, light trail, strong outline, all cap	narrative driven, manga, Japanese, cartoon drawing, graphic art	regulated layout, posh, promoting, triangle composition, branding package	mixed media, layered composition, watercolor painting, handmade artwork painting, illustrative	color splash, marker drawing, expressionist, word sound effects, subtle colors
ALADIN-NST (Fused)	dark image, chiaroscuro, black and red, red, dark picture	documentary, vigorous, past, documentary shot, dark contrast	trustful, housing, posh, housing architecture render, solarpunk	colorful drawing, abstract art, abstract artwork, nonobjective, cubism	storyboarding, interpretive, panel, story, storyboarding
ALADIN-NST	flame, spark, dark vibe, explosion, dark space	contrasted, high contrast, suburban, documentary shot, documentary	glass, trustful, architectural landscape, pentagonal, pointy	soft, soft color, soft and bright color, feminine, soft colors	comic book art, black and white art, comic art, comic, doodle art