PandA: Unsupervised Learning of Parts and Appearances in the Feature Maps of GANs

James Oldfield [1] Christos Tzelepis [1] Yannis Panagakis [2] Mihalis A. Nicolaou [3] Ioannis Patras [1]

[1]Queen Mary University of London [2]University of Athens [3]The Cyprus Institute

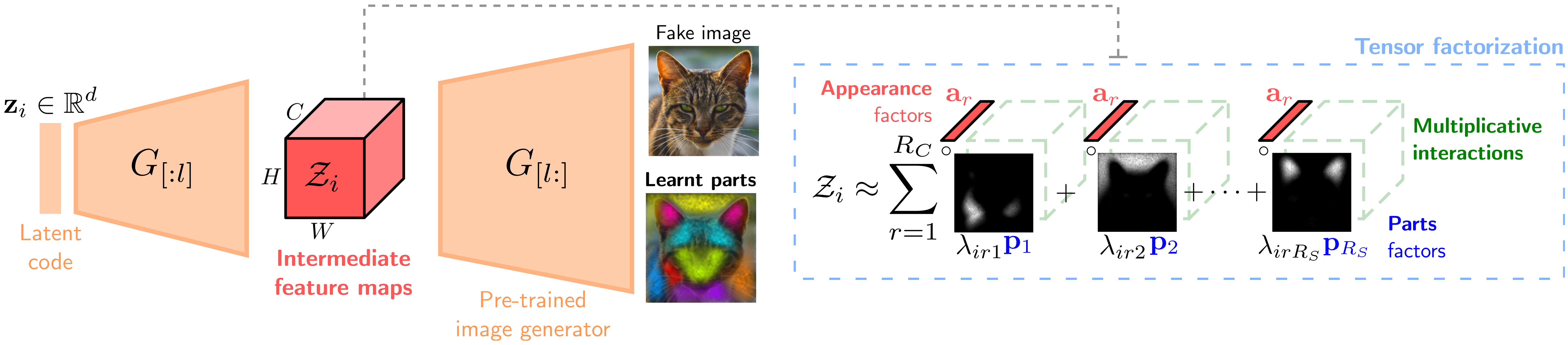


Figure 1. Overview of our unsupervised factorisation of a dataset of synthetic samples' feature maps: the parts (non-negative) and appearance factors are learnt in the spatial and channel modes respectively; combinations of which are combined with the outer product and sample-specific coefficients.

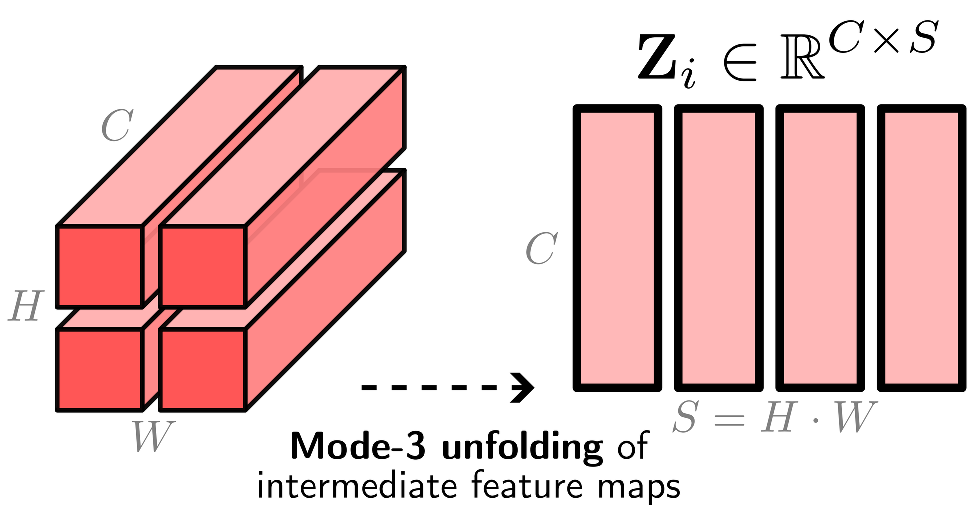
Overview

We propose an unsupervised factorisation of a dataset of pretrained generator's intermediate feature maps. This provides an intuitive separation into representations of an image's parts and appearances. The learnt semantic factors allow for:

- **Local image editing**: precise *pixel-level* control not facilitated by the SOTA.
- **Context-aware object removal**: a single appearance factor removes objects in a scene.
- **Concept localization**: the appearance factors localize semantic concepts in the image, such as the sky, skin, or background.

Method

Let $\mathbf{Z}_i \in \mathbb{R}^{C\times S}$ be sample i's feature maps with their C-dimensional channel fibers stacked along the columns:



We write each sample's feature maps as its own combination of shared appearance and non-negative parts factors $\mathbf{A} \in \mathbb{R}^{C\times R_C}$ and $\mathbf{P} \in \mathbb{R}^{S\times R_S}$ respectively:

$$\mathbf{Z}_i = \mathbf{A}\mathbf{\Lambda}_i\mathbf{P}^\top$$

$$= \underbrace{\begin{bmatrix} | & & | \\ \mathbf{a}_1 & \cdots & \mathbf{a}_{R_C} \\ | & & | \end{bmatrix}}_{\text{Appearance}} \underbrace{\begin{bmatrix} \lambda_{i11} & \lambda_{i12} & \cdots \\ \vdots & \ddots & \\ \lambda_{iR_C1} & \cdots & \lambda_{iR_CR_S} \end{bmatrix}}_{\text{Sample } i\text{'s coefficients}} \underbrace{\begin{bmatrix} - & \mathbf{p}_1^\top & - \\ & \vdots & \\ - & \mathbf{p}_{R_S}^\top & - \end{bmatrix}}_{\text{Parts}}$$

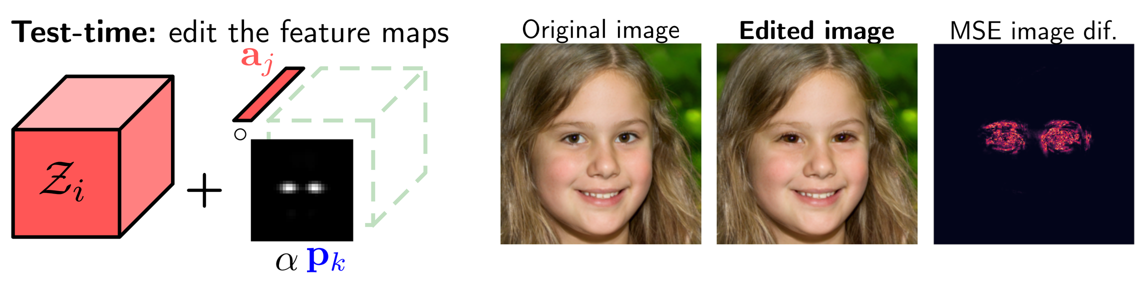
The global factor matrices are learnt by formulating and solving the following constrained optimisation problem:

$$\min_{\mathbf{A},\mathbf{P}} \sum_{i=1}^{N} ||\mathbf{Z}_i - \mathbf{A}\underbrace{\left(\mathbf{A}^\top\mathbf{Z}_i\mathbf{P}\right)}_{\mathbf{\Lambda}_i}\mathbf{P}^\top||_F^2 \quad \text{s.t. } \mathbf{P} \geq \mathbf{0}.$$

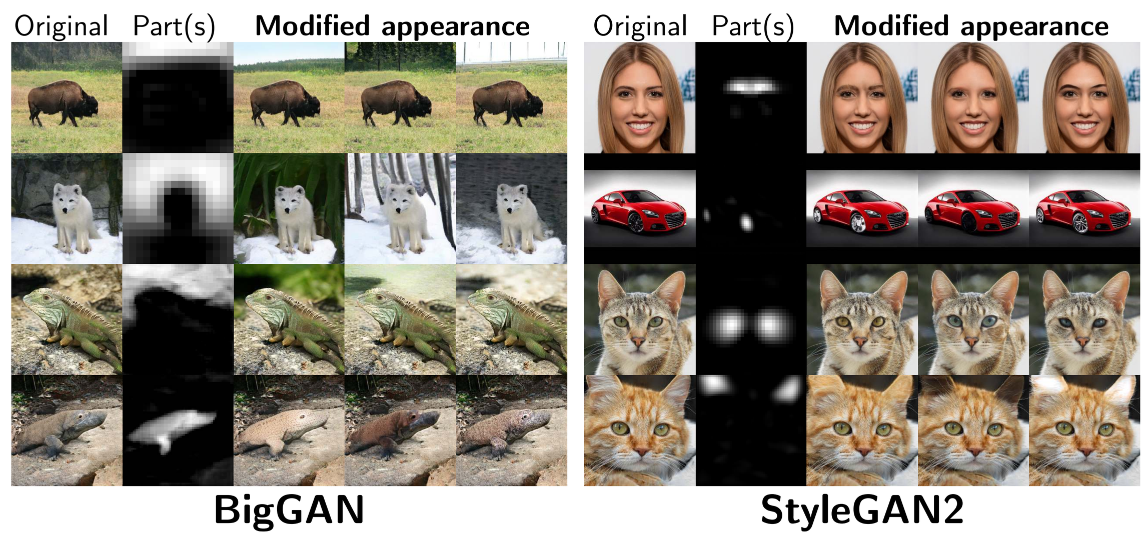
Refinement: if desired, one can subsequently optimise for sample-specific parts factors $\tilde{\mathbf{P}}_i$ for particular images/datasets lacking alignment.

Local image editing

To locally modify an image i at region k with the j^{th} appearance with desired magnitude $\alpha \in \mathbb{R}$, we compute the forward pass from layer l onwards in the generator with $G_{[l:]}\left(\mathbf{Z}_i + \alpha\mathbf{a}_j\mathbf{p}_k^\top\right)$.

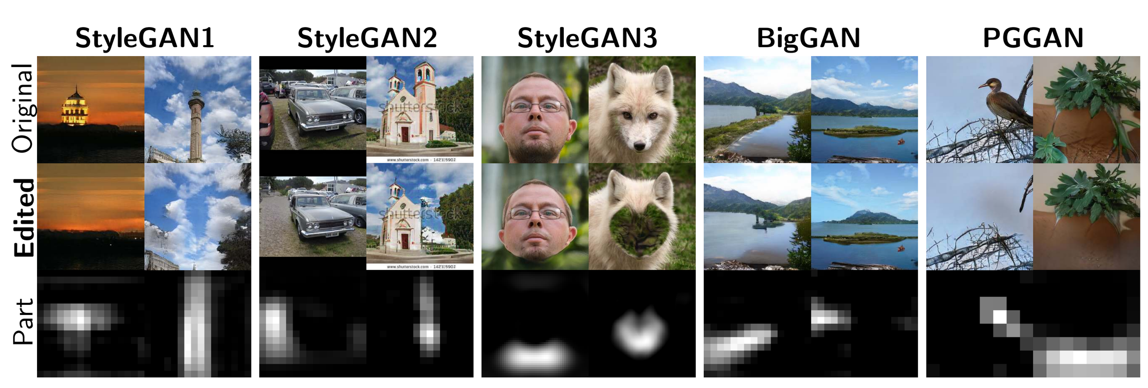


Unlike the SOTA [4, 5, 3], the proposed method **requires neither manually defined ROIs, nor semantic masks**, and is orders of magnitude faster to train.



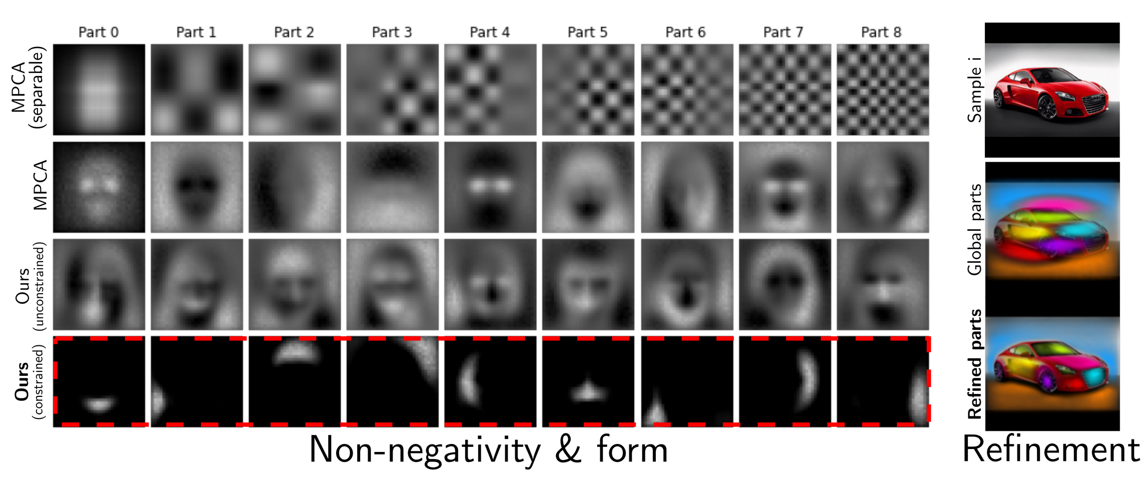
Context-aware object removal

We find the decomposition frequently learns an appearance factor $\mathbf{a}_b$ that controls a high-level 'background' concept in all 5 generator architectures studied.



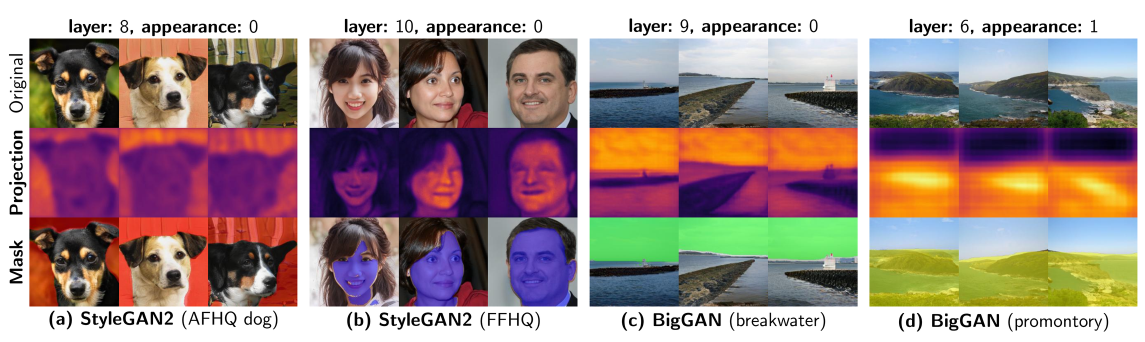
Using $\mathbf{a}_b$ (learnt for a particular generator and dataset) one can remove objects from an image by simply updating the feature maps with $\alpha\mathbf{a}_b\mathbf{p}_k^\top$ as above.

Ablations



Concept localization

- The columns of $\mathbf{A}\mathbf{A}^\top\mathbf{Z}_i = \mathbf{Z}_i \in \mathbb{R}^{C\times S}$ contain the activations at each of the S spatial positions in sample i's feature maps.
- $\mathbf{A}^\top\mathbf{Z}_i \in \mathbb{R}^{R_C\times S}$, viewed as a change of basis (when $R_C = C$), tells us 'how much' each of the appearance factors is present at the S spatial positions. This interpretation readily localizes the learnt concepts in the images:



Quantitative results

Quantifying local image editing precision: the norm of the difference between the edited and original images outside the ROI, divided by the same quantity inside the ROI:

Table 1. ROIR (↓) of for 10k FFHQ samples per local edit.

	Eyes	Nose	Open mouth	Smile
GANSpace [1]	2.80±1.22	4.89±2.11	3.25±1.33	2.44±0.89
SeFa [2]	5.01±1.90	6.89±3.04	3.45±1.12	5.04±2.22
StyleSpace [3]	1.26±0.70	1.70±0.82	1.24±0.44	2.06±1.62
LowRankGAN [4]	1.78±0.59	5.07±2.06	1.82±0.60	2.31±0.76
ReSeFa [5]	2.21±0.85	2.92±1.29	1.69±0.65	1.87±0.75
Ours	1.04±0.33	1.17±0.44	1.04±0.39	1.05±0.38

References

[1] Erik Härkönen et al. "GANSpace: Discovering Interpretable GAN Controls". In: *NeurIPS*. 2020.

[2] Yujun Shen and Bolei Zhou. "Closed-Form Factorization of Latent Semantics in GANs". In: *CVPR*. 2021.

[3] Zongze Wu et al. "StyleSpace analysis: Disentangled controls for stylegan image generation". In: *CVPR*. 2021.

[4] Jiapeng Zhu et al. "Low-Rank Subspaces in GANs". In: *NeurIPS*. 2021.

[5] Jiapeng Zhu et al. "Region-Based Semantic Factorization in GANs". In: *ICML*. 2022.

SCAN ME for project page