Exploring Image Representation with Decoupled Classical Visual Descriptors

Chenyuan Qu[1]
cxq134@student.bham.ac.uk
Hao Chen[1,2]
hc666@cam.ac.uk
Jianbo Jiao[1]
j.jiao@bham.ac.uk

[1] University of Birmingham
Birmingham, UK
[2] University of Cambridge
Cambridge, UK

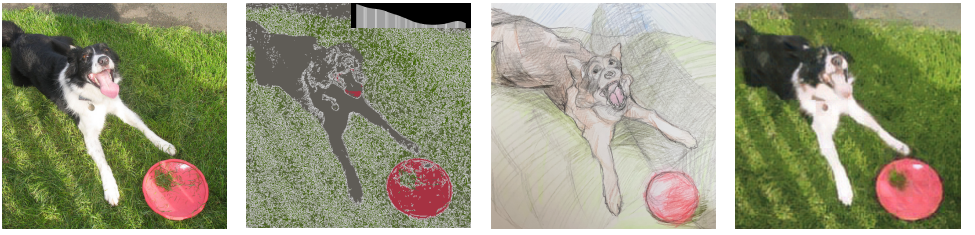
Abstract

Decomposing visual components remains a fundamental challenge in computer vision. While deep learning has achieved remarkable progress across image understanding tasks, its internal representations are often opaque, making it difficult to interpret how visual information is processed. In contrast, classical visual descriptors (e.g. edge, colour, and intensity distribution) have long been fundamental to image analysis and remain intuitively understandable to humans. Inspired by this, we present VisualSplit, a new paradigm designed to explore image decomposition by explicitly leveraging decoupled classical visual descriptors. Specifically, it integrates multiple traditional visual descriptors as independent yet informative inputs, extracting meaningful features through a reconstruction-based pretraining process. By explicitly decomposing visual attributes, our method inherently facilitates effective attribute control in various advanced visual tasks, including image generation and editing, extending beyond conventional classification and segmentation, suggesting the effectiveness of this new learning approach towards visual understanding. The code and models will be made publicly available.

1 Introduction

Computer vision, and more recently deep learning, has continuously pursued effective *decomposition* of visual information, to extract semantically meaningful visual components directly from input images, thereby achieving a deeper and clearer understanding of their inherent content. In classical computer vision, feature extraction relied heavily on predefined methods such as dimensionality reduction [21], low-level feature descriptors [13], and handcrafted feature extractors [7, 20]. While these traditional methods offer strong deterministic properties, they are limited in adaptability and scalability, as they are tailored to specific tasks without the ability to generalise across diverse scenarios.

In contrast, contemporary deep learning strategies approach visual decomposition through learning-based paradigms such as supervised learning [18, 32] and self-supervised pretext learning [39]. For instance, RetinexNet [37] decomposes images into reflectance and illumination components, enabling controlled lighting adjustments while preserving colour. Similarly, HiSD [18] effectively isolates essential human facial attributes, facilitating fine-grained



(a) Original Image (b) Combined Features (c) Human Depicted (d) VisualSplit(Ours)

Figure 1: **Qualitative overview of the key idea.** (a) Original image. (b) Visualisation of a combination of threefold visual descriptors (grey-level histogram displayed in the top-right). (c) Image completion by a human artist using (looking at) only the combined descriptor set. (d) Recovery from the proposed VisualSplit, using only the same incomplete descriptor set.

control over expressions and facial feature modifications. More broadly, the split-brain autoencoder [39] demonstrates image decomposition by partitioning visual inputs according to the LAB colour channels, enhancing feature extraction and understanding.

Main idea: Despite extensive studies on classical visual descriptors from traditional computer vision with well-established deterministic properties, their potential in the context of deep learning models has been largely overlooked. In this paper, we are interested in the question: *"If classical visual descriptors can be used to decompose visual information in learning-based frameworks?"*. From an artistic perspective, images are characterised by stylistic descriptors such as *line, colour, and value* [25], which closely align with the key components for computational visual understanding and analysis [3]. Inspired by these parallels, we seek to leverage conventional computer vision algorithms to extract visual descriptors—specifically *edge, colour segmentation map and grey-level histogram* (Section 3). Each descriptor independently and sparsely captures a distinct aspect of the visual input. Notably, the combined descriptor depiction offers only partial information of the image (Figure 1b).

Although these descriptors are highly abstract and compressed compared to the original image, humans can still figure out the main underlying semantic content. To this end, we present a case study in which an artist was asked to depict an image given such "abstract information". Interestingly, the artist was able to reconstruct an image (Figure 1c) very similar to the original one (Figure 1a). We attribute this ability to the human capacity for understanding incomplete visual concepts. Motivated by this, we seek to explore if such ability can be modelled by representation learning. To achieve that, we introduce VisualSplit, a mask-free framework that follows such visual understanding process by learning to integrate (only) these classical descriptors to reconstruct the underlying image (Figure 1d), which achieves high representation quality while bringing extra controllability (details in Section 5).

Remark: Although VisualSplit is implemented in a self-supervised manner, the primary focus of this work is on the study of decoupled visual representations rather than conventional evaluation pipeline in self-supervised representation learning literature. Our goal is not to surpass existing representation learning methods on those benchmarks, but rather aiming at answering the question we asked above. In summary, our key contributions are as follows:

- We revisit classical visual descriptors and introduce a new learning paradigm – *VisualSplit*, to explore the representation learning capacity using such simple descriptors.
- Aligning closely with human perceptual understanding, the proposed approach highlights the potential of the overlooked classical while effective visual descriptors.
- Extensive experimental analysis on low-level (4.4) and high-level (4.3) vision tasks,

covering various applications, validates the effectiveness of our *VisualSplit* learning approach. It shows that precise, independent, and intuitive manipulation of image attributes (such as geometry, colour, and illumination) can be achieved (4.5 and 4.6).

2 Related Work

Visual descriptors. In the field of computer vision, visual descriptors are widely used to represent image content. In the classic computer vision area, traditional visual descriptors are manually designed to capture specific image properties, including edge maps [27] [5], colour histograms [28], Scale-Invariant Feature Transform (SIFT) [20], and Histogram of Oriented Gradients (HOG) [7]. Despite their usefulness and deterministic nature, traditional descriptors suffer from rigidity, sensitivity to variations such as lighting and perspective changes, and limitations inherent in manual feature engineering. In recent years, deep learning approaches have relied mainly on learning-based methods, particularly convolutional neural networks (CNNs) and Vision Transformers (ViTs). These approaches automatically learn rich feature representations from extensive datasets, significantly improving robustness and descriptive power. [26] [4] [9] [29] However, learning-based methods are not easily interpretable and usually mix different attributes and patterns.

Visual decoupling and disentanglement. Decoupling visual features focuses on eliminating dependencies between components, whereas disentanglement emphasizes the separation of distinct underlying factors that correspond to meaningful and independent concepts. These principles are of great importance in various applications such as visual generation [14, 34], model generalization [6, 11], and model explainability [36, 39]. Given their conceptual overlap, we examine both lines of research in this discussion. The SplitBrain model [39] decouples representations by independently processing the LAB channels, enabling the isolation of distinct channel contributions to the overall task. Li *et al*. [17] propose decoupling an image into low-frequency and high-frequency elements, which correspond to body features and edge features, respectively, with a stronger focus on semantic segmentation tasks. Yang *et al*. [34] develop a causal disentanglement approach to align latent factors with the semantics of interest. BayeSeg [11] disentangles domain-invariant features to enable generalization to unseen data.

3 Methodology

As illustrated in Figure 2, the proposed VisualSplit approach learns decoupled representations by reconstructing the original visual signal from isolated, incomplete visual descriptors. VisualSplit comprises three key components: descriptor extraction, a multi-modal encoder, and a lightweight decoder for image reconstruction.

Descriptor extraction. The VisualSplit leverages traditional computer vision algorithms to extract three deterministic and human-understandable descriptors: *edge maps*, *colour segmentation maps*, and *grey-level histograms*, denoted as d_e, d_c and d_g, respectively.

To extract the desired descriptors, we first transform the image x from the RGB domain into the LAB domain, denoted as x_{LAB}. Next, we apply the Convolutional Sobel operator [13] to the L-channel of the image ($d_e = \text{Sobel}(x_{\text{L}})$), designed to extract of edge descriptors. Simultaneously, we use the L-channel to generate the smooth grey-level histogram ($d_g = \text{Hist}(x_{\text{L}})$) through a Gaussian Kernel with 100 bins providing valuable insights into

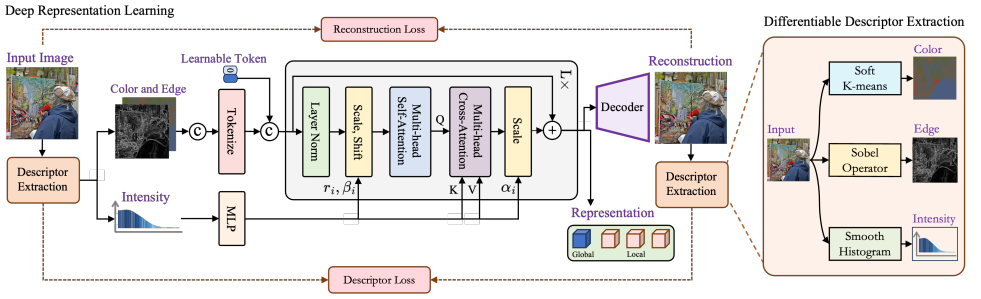


Figure 2: **VisualSplit framework.** Initially, the RGB image is processed through the traditional *Descriptor Extraction* module to obtain segmented colour, edge, and intensity histograms. The colour and edge data will be patchtified and fed into the transformer, whereas the intensity information will serve as a condition for Cross-attention and to learn scale and shift parameters (r, β) for AdaLN-Zero. The loss is calculated on an image level as well as on the traditional descriptor level. It is worth mentioning that the acquired model can output both global and local representations, which can be applied to downstream visual tasks.

the intensity distribution. For colour segmentation, we apply the Soft K-Means clustering algorithm, which partitions the image into K distinct clusters ($d_c = \text{Cluster}(x_{\text{AB}})$) based on colour values. Noteworthy, these descriptors are differential approach (Supp. Section 2.2).

Existing self-supervised representation learning methods are typically built upon the principle of missing information prediction, while the common approach is explicit masking strategies. In contrast, our method employs classical visual descriptors, which inherently produce sparse and abstract representations of the original image. These descriptors can be regarded as a form of information absence, therefore, additional artificial patch masking is unnecessary. Moreover, our approach effectively preserves and reinforces globally and locally meaningful visual structures within the learned representations.

Multi-model encoder. Visual descriptors describe images either locally or globally. In our case, edge and colour segmentation descriptors primarily capture localised pixel-level information, whereas the grey-level histogram captures global information. Given this distinction, it is essential to employ a multi-modal encoder capable of considering each type of descriptor simultaneously, while also mapping local and global information into latent spaces, ensuring a comprehensive representation.

Building on this observation, we construct a multi-modal encoder based on ViT [30], which operates on local patches. To accommodate the distinct roles of local and global descriptors, we treat local descriptors as patch-wise inputs for backbone learning, while incorporating global descriptors, such as histograms, to provide global conditioning.

The conditioning process consists of two key components: AdaLN-Zero [23] and a Multi-head Cross-attention Layer, each operating within separate blocks. AdaLN-Zero learns parameters that scale and shift intermediate features derived from an MLP, which takes d_g as input. Simultaneously, the multi-head cross-attention layer utilises d_g as the key-value pair, allowing the model to integrate relevant information from a global perspective.

To further enhance the learning of global representation, we introduce additional learnable embedded patches that represent the global state at the output followed by the design of ViT [30], while the remaining outputs correspond to local representations.

Decoder. Our model employs a more lightweight decoder specifically designed for the pre-training phase. The primary objective is to ensure that the encoder learns more complex

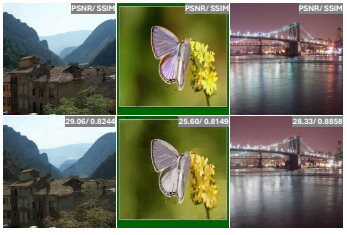


Figure 3: **Visualisation of pre-train reconstruction task.** Top: original images, bottom: our restoration results.

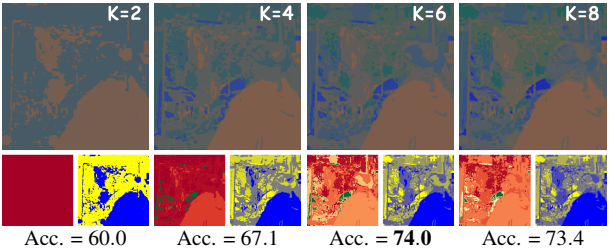


Figure 4: **Clustering visualisation with K clusters.** Each set shows the colour-clustered LAB image (top), alongside the isolated A-channel and B-channel (bottom left/right). Acc. indicates the classification accuracy of linear probing.

representation than the decoder. To achieve this, the decoder is intentionally kept much shallower than the encoder, a design choice that also accelerates the training process. We discussed it further in Supp. Sec. 2.3.

Pretraining objectives. The goal of our pretraining is to reconstruct the original image. Therefore, the primary objective during this phase is to minimize the discrepancy between the reconstructed image $\hat{x}_{\text{RGB}}$ and the original image x_{RGB} in the pixel space. To achieve this, we compute the mean squared error (MSE) between the two images. Additionally, we incorporate the LPIPS loss function [40] to assess reconstruction quality in latent space, providing a more comprehensive and perceptually aligned evaluation.

Furthermore, the encoder is required to effectively capture input visual descriptors and map them to latent representations. To ensure its capability to do so, we introduce a descriptor consistency loss. By applying the same processing operations to the reconstructed image in pixel space as to the input, we apply visual descriptors on the reconstructed image. The output of these descriptors are then compared with the original ones to compute the descriptor consistency loss, consisting of three terms:

$$L_{\text{e}} = \|d_e - \hat{d}_e\|_1^1, \quad L_{\text{g}} = \frac{1}{N}\sum_i^N \frac{(d_g^i - \hat{d}_g^i)^2}{d_g^i + \hat{d}_g^i}, \quad L_{\text{c}} = \|d_c - \hat{d}_c\|_1^1, \tag{1}$$

where d_e, d_g and d_c represent edges, grey-level histogram, and colour segmentation map, respectively. Ablation studies on objective functions are shown in Supp. Sec. 4.

4 Experiments

4.1 Implementation Details and Benchmark

Our method is compatible with ViT models of any size; for simplicity, we adopt ViT-Base as the backbone in our experiments. The model is self-supervised and pre-trained on ImageNet-1K [8]. Our implementation mainly follows ViTMAE [17], using the AdamW [19] optimizer with a cosine decay learning rate schedule and an initial learning rate of 1.5×10^{-4}. All images are resized and center-cropped to 224×224. Full implementation and training details are provided in the supplementary material.

Figure 5: Reconstruction results with grey-level histograms at different brightness levels.

Figure 6: Reconstruction with different colour segmentation map inputs extracted from the image after different colour balance processing, where f is the enhancement factor.

4.2 Performance on the Pre-training Task

The pretraining task is formulated as image reconstruction. Figure 3 presents qualitative results along with PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index Measure) scores [31]. The results demonstrate that the model can effectively reconstruct images using only isolated and incomplete descriptors.

Colour segmentation clusters. Choosing an appropriate number of clusters is critical for colour segmentation. Too few clusters lead to excessive information loss, degrading performance; too many make the segmented map resemble the original image, oversimplifying the task and limiting the model's ability to learn meaningful representations. Figure 4 illustrates the visual impact of varying cluster counts and their effect on linear probing performance. The model performs best when the number of clusters is set to 6.

Input independence. We evaluate how different inputs independently influence the model's output. Ideally, modifying one input should affect only specific attributes of the output while leaving others largely unchanged.

We first vary the grey-level histogram, which captures image brightness. We use the SIDD dataset [1], which offers images under low, normal, and high exposure, we extract all inputs from the normal image. For other exposures, only the histogram is updated, keeping edge and colour descriptors fixed. Figure 5 shows the qualitative results.

To assess the effect of colour segmentation, we use synthetic examples due to the difficulty of sourcing real images with identical edges and brightness but different colours. Starting with an image from the Flickr dataset [37], we extract all descriptors. We then alter its colour balance to generate variations and extract new colour segmentation maps, which are combined with the original edge and histogram inputs. Visual results are shown in Figure 6.

4.3 Representation Analysis on Classification

Setting. To evaluate the performance of the representations of different models, we conduct classification experiments with fine-tuning and linear probing. In this experiment, we extract representations from the pre-trained encoder and a lightweight header will be added for fine-tuning and linear probing. An additional learnable embedded patch is added to the model at the input of the encoder to serve its state at the output as a global representation. The full implementation detail is also shown in the supplementary material.

To validate the efficacy of our proposed methodology, we conducted comparisons with previous self-supervised methods, Split-brain Autoencoder [39], MAE [12], and PeCo [10]. To ensure fair comparisons, the backbone of all the models is ViT-base architecture. We replace the Split-brain Autoencoder encoder and decoder with ViT-base architecture, as the original implementation is CNN-based. Meanwhile, we compared the performance with our implementation without pre-training to validate the effectiveness of the pre-training method.

Table 1: Classification generalisation in Accuracy (%).

Method	Classification (Acc.%) Finetuning	Linear Probing
scratch, our impl.	82.0	0.1
Split-brain Auto [39]	82.3	36.4
MAE [13]	83.6	72.4
PeCo [10]	**84.5**	51.7
VisualSplit (Ours)	83.5	**74.0**

Table 2: Transfer learning for classification and segmentation.

Method	Classification (Acc.%) Places dataset	Segmentation (mIoU) ADE20K dataset
scratch, our impl.	79.9	47.7
Split-brain Auto [39]	81.4	44.4
MAE [13]	82.5	48.4
PeCo [10]	82.4	48.7
VisualSplit (Ours)	**82.7**	**49.7**

Results. In Table 1, we compare the quantitative results of different methods of fine-tuning and linear probing. For fine-tuning, the performance differences among all methods are not significant; our method is comparable to ViTMAE [13] and slightly inferior to PeCo. However, our model performs better than the other models in linear probing, which indicates that our model can extract the useful underlying representations from images better.

4.4 Representation Analysis on Transfer Learning

To further evaluate the performance of the extracted representation, we used the classification task pretrained encoder from Section 4.3 to assess transfer learning in downstream tasks.

First, we tested its classification generalisation on other datasets. Additionally, we conducted a Semantic Segmentation experiment on ADE20K [42], leveraging the segmentation head from Segformer [33], which is a lightweight decoder composed solely of MLPs. The detailed structure is provided in the supplementary material.

Table 2 presents transfer learning results on classification and segmentation using the Places [41] and ADE20K [42] datasets, respectively. Our model achieves the best transfer performance, *e.g.* 49.7 mIoU in semantic segmentation, outperforming MAE and PeCo by +1.3 and +1.0, respectively. This advantage may arise from pretraining on visual descriptors rather than raw images, encouraging the model to learn underlying structures instead of dataset-specific biases.

4.5 Representation for Visual Restoration

Setting. In this section, we utilise image restoration tasks to intuitively demonstrate the effectiveness of our learned representations. In this task, we combine with pre-trained Stable Diffusion 1.5 [24]. Our learnt representation encompasses both global and local representations. We employ the global representation combined with the original text embedding, inspired by IP-Adapter [35], providing an overall condition for the generated image. Meanwhile, the local representations are integrated via ControlNet [38] to guide the structural generation within the UNet model, enabling precise control in the image regions.

To enable an effective comparison with existing guidance methods, ControlNet [38], T2I-Adapter [22], ControlNet++ [16], we use the edge map and segmented colour map directly as inputs for these baseline methods. We also input the grey-level histogram directly through IP-Adapter. However, we found that using only the grey-level histogram in IP-Adapter might limit the model ability during genera-

Table 3: Image restoration results.

Method	Prompt	PSNR	SSIM
ControlNet [38]	w/o	17.34	0.6374
	w/	16.52	0.6051
T2I Adapter [22]	w/o	17.69	0.5421
	w/	17.30	0.5459
ControlNet++ [16]	w/o	19.94	0.6549
	w/	19.50	0.6399
VisualSplit (Ours)	w/o	**26.56**	**0.8664**

Figure 7: Comparison of image restoration results using Stable Diffusion 1.5 with various guidance methods, which includes ControlNet [58], T2I-Adapter [22], ControlNet++ [16], and our proposed method, with the ground truth shown in the final column. The condition inputs are edge, segmented colour map, and grey-level histogram inputs, shown in the first column. Each method, except for our method, displays results with and without prompts, where prompts are generated by BLIP [15].

tion. For further comparison, we additionally employ
BLIP [15] to generate a text prompt, which is then fed into the text encoder.

Results. As shown in Table 3 and Figure 7, our method reconstructs high-quality images, preserving both global features and local details better than baseline approaches. This highlights the advantage of our structured representations over raw inputs, improving model interpretability and generalizability. Moreover, our approach effectively manages multiple representation controls, overcoming the limitations of traditional methods.

4.6 Representation for Visual Editing

Our approach benefits from separating visual descriptors in advance. In image editing task, it allows us to edit these descriptors with much greater ease than directly manipulating and controlling the image. In this section, we demonstrate how simple adjustments to the segmented colour map and grey-level histogram enable controllable image editing without any additional training or modification on models. We followed the setting of Section 4.5, and only edited the original segmented colour map to edit the colour regions of the generated image and keep the other inputs the same. The region of the colour map will only be recoloured to ensure it will not conflict with the edge.

As illustrated in Figure 8, modifying the segmented colour map provides a straightforward way of directing the model to edit the original image with precise control. Notably, our method maintains image harmony while performing controlled editing. We also have modified the grey-level histogram for image editing to adjust the brightness of the output, which is shown in Supp. Sec. 5.2. We further conducted a survey based on human perception for the qualitative analysis of image editing tasks in Supp. Sec. 5.1. The result shows that our method significantly outperforms the others in all attributes.

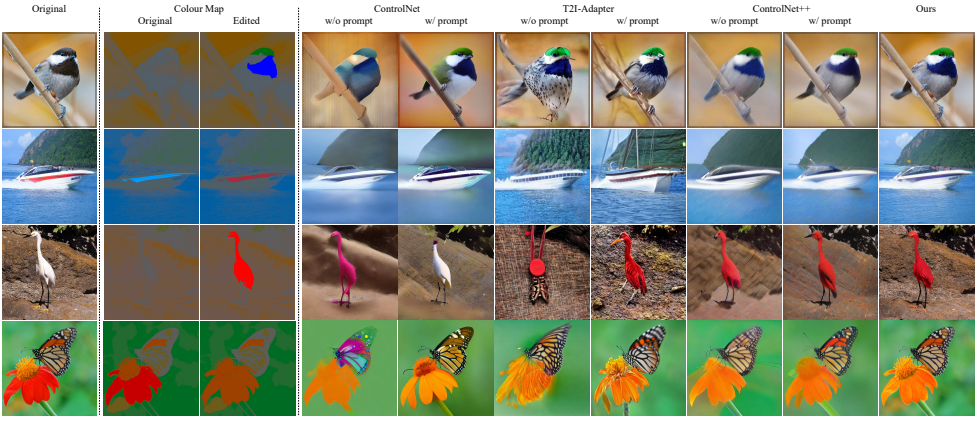


Figure 8: Comparison of colour editing results using modified segmented colour maps as input to Stable Diffusion 1.5, guided by ControlNet [33], T2I-Adapter [22], ControlNet++ [16], and our method. The first column shows original images, followed by original and edited colour maps. For each baseline, results are shown with and without prompts generated by BLIP using edited colour keywords. Our method requires **no prompts**.

5 Discussion

Mask-free training. Unlike random pixel dropout, our mask-free approach eliminates the need for masking ratio tuning. From an information-theoretic perspective, it leverages only 17% of the RGB signal, less than Split-Brain Autoencoder (50%), MAE (25%), and PeCo (60%), which encourages model to learn better representation [2]. See Supp. Sec. 3. As shown in Section 4.3, our method slightly underperforms baselines after fine-tuning but notably outperforms them under linear probing, suggesting more explicit representations. It also generalises across datasets and tasks (Section 4.4), while in generation tasks (Section 4.5). The learnt representation adapts seamlessly to other models.

Decouple controllable attributes. Section 4 demonstrates VisualSplit's ability to isolate and control specific attributes (*e.g.* geometry, colour, and illumination) via decoupled visual inputs. In Section 4.2, we show that these inputs remain independent of the generated outputs during pre-training. Section 4.6 further illustrates that our descriptors can independently and effectively guide the generation process without requiring any model modification, enabling precise control over individual attributes without altering the whole.

While our method integrates classical visual descriptors and performs well across tasks, its limitations and potential directions for future work are discussed in Supp. Sec. 7 and 6.

6 Conclusion

In this work, we investigated the challenge of decoupling the visual content by leveraging classic visual descriptors through learning-based paradigms, which have been largely overlooked in deep models. Inspired by the human ability to infer and reconstruct visual content from incomplete descriptor combinations, we explore whether deep models can emulate this capacity and propose VisualSplit, a new visual decoupling paradigm. Our experiments demonstrate that the learned representations not only excel in low-level tasks such as classification and segmentation but also effectively guide high-quality image generation and editing

through decoupled descriptor inputs. Furthermore, this approach opens new possibilities for advanced image manipulation and encourages further research into decoupling-based models that integrate conventional visual descriptors. We hope this work provides a fresh perspective on visual decoupling and inspires future research on its extensive applications across various visual tasks.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700, 2018.

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[3] Roger D Boyle and Richard C Thomas. *Computer vision: A first course*. Blackwell Scientific Publications, Ltd., 1988.

[4] Andrei Bursuc, Giorgos Tolias, and Hervé Jégou. Kernel local descriptors with implicit rotation matching. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 595–598, 2015.

[5] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[6] Hao Chen, Hongrun Zhang, U Wang Chan, Rui Yin, Xiaofei Wang, and Chao Li. Domain game: Disentangle anatomical feature for single domain generalized segmentation. In *International Workshop on Computational Mathematics Modeling in Cancer Analysis*, pages 41–51. Springer, 2024.

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Jingming Dong and Stefano Soatto. Domain-size pooling in local descriptors: Dsp-sift. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5097–5106, 2015.

[10] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu, and Baining Guo. Peco: Perceptual codebook for bert pre-training of vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 552–560, 2023.

[11] Shangqi Gao, Hangqi Zhou, Yibo Gao, and Xiahai Zhuang. Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. *arXiv preprint arXiv:2303.01710*, 2023.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[13] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23 (2):358–367, 1988.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[16] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++ : Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pages 129–147. Springer, 2025.

[17] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 435–452. Springer, 2020.

[18] Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, and Rongrong Ji. Image-to-image translation via hierarchical style disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8639–8648, 2021.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[20] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[21] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

[22] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.

[23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[25] Marjo Räsänen. *Interpreting Art Through Visual Narratives*, volume 2, pages 183–195. 01 2003. ISBN 978-1-4020-1637-0. doi: 10.1007/978-94-010-0043-7_13.

[26] Johannes L Schonberger, Hans Hardmeier, Torsten Sattler, and Marc Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1482–1491, 2017.

[27] Irwin Sobel, Gary Feldman, et al. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, 1968:271–272, 1968.

[28] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

[29] Kai Tian, Shuigeng Zhou, and Jihong Guan. Deepcluster: A general clustering framework based on deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II 17*, pages 809–825. Springer, 2017.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[32] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement.

[33] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[34] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9593–9602, 2021.

[35] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

[36] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2327–2336, 2019.

[37] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[38] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[39] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1058–1067, 2017.

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[41] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[42] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.