Learning Invariant Inter-pixel Correlations for Superpixel Generation

Sen Xu[1,2], **Shikui Wei**[*1,2], **Tao Ruan**[3], **Lixin Liao**[4]

[1]Institute of Information Science, Beijing Jiaotong University
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology
[3]Frontiers Science Center for Smart High-speed Railway System, Beijing Jiaotong University
[4] DaoAI Robotics Inc.
{senxu1, shkwei, ruantao} @bjtu.edu.cn, liaolixin@163.com

Abstract

Deep superpixel algorithms have made remarkable strides by substituting hand-crafted features with learnable ones. Nevertheless, we observe that existing deep superpixel methods, serving as mid-level representation operations, remain sensitive to the statistical properties (e.g., color distribution, high-level semantics) embedded within the training dataset. Consequently, learnable features exhibit constrained discriminative capability, resulting in unsatisfactory pixel grouping performance, particularly in untrainable application scenarios. To address this issue, we propose the **C**ontent **D**isentangle **S**uperpixel (**CDS**) algorithm to selectively separate the invariant inter-pixel correlations and statistical properties, i.e., style noise. Specifically, We first construct auxiliary modalities that are homologous to the original RGB image but have substantial stylistic variations. Then, driven by mutual information, we propose the local-grid correlation alignment across modalities to reduce the distribution discrepancy of adaptively selected features and learn invariant inter-pixel correlations. Afterwards, we perform global-style mutual information minimization to enforce the separation of invariant content and train data styles. The experimental results on four benchmark datasets demonstrate the superiority of our approach to existing state-of-the-art methods, regarding boundary adherence, generalization, and efficiency. Code and pre-trained model are available at https://github.com/rookiie/CDSpixel.

Introduction

Superpixel segmentation (Achanta et al. 2012; Liu et al. 2011; Achanta and Susstrunk 2017; Yang et al. 2020; Wang et al. 2021) divides an image into compact and contiguous regions of pixels based on specific criteria such as color similarity, texture, or brightness. Compared to pixel-level processing, superpixels significantly reduce the number of image primitives and preserve the structural information, thus improving the efficiency and accuracy of many downstream tasks, i.e., semantic segmentation (He et al. 2015; Zhu et al. 2014; Kwak, Hong, and Han 2017; Gadde et al. 2016), stereo matching (Yang et al. 2020; Birchfield and Tomasi 1999; Wang and Zheng 2008), self-supervised pretraining (Sautier et al. 2022), image classification (Zhao, Zhu, and Feng 2022; Guo et al. 2018), etc. The usual practice of traditional superpixel algorithms is to initialize a regular grid of superpixels and iteratively adjust the association of pixels and neighboring superpixels. Afterward, deep learning algorithms leverage image features learned by neural networks to replace hand-crafted features, i.e., XYlab used by traditional methods, significantly improving the performance of superpixel algorithms.

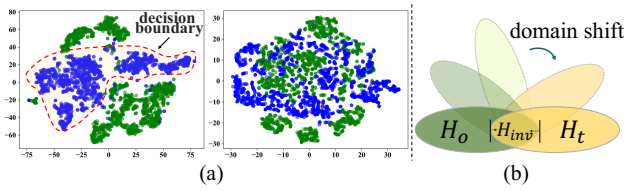


Figure 1: Motivation. (a) Visualization of the t-SNE distributions on the BSDS dataset. From left to right are the baseline and our CDS. After applying color inversion, the feature distribution of baseline displays a noticeable decision boundary. In contrast, the feature distribution extracted by CDS is more compact (i.e.,[-80, +80] vs [-40, +40]) and indivisible. (b) Gradually modifying the stylistic information of both auxiliary (H_t) and original data (H_o) enhances the purity of the shared invariant inter-pixel correlations (H_{inv}).

However, although learnable features have been proven effective, they also introduce new challenges. Superpixel segmentation, as a mid-level image representation task, needs to adapt to a variety of open-world scenarios. Traditional superpixel algorithms employ independent online processes (e.g., clustering or graph-cut), which remain unaffected by inter-instance influences. In contrast, deep superpixel algorithms carry the potential risk of learning the unique data distribution present in the training set. To validate this potential risk, we conducted a straightforward experimental verification in Fig. 1(a). Given that this concern reflects the algorithm's generalization ability, we opted for SCN (Yang et al. 2020), a deep superpixel algorithm renowned for its solid generalization, for experimental validation. We applied an inversion operation (subtracting each pixel value from 255) to simulate variations in the data distribution within the test set. While the inter-pixel correla-

*Corresponding author

tions of identical images should remain constant through the linear transformation of RGB values during superpixel generation, the features extracted by the baseline algorithm exhibit a discernible decision boundary. Hence, learnable superpixels not only encompass a pure representation of inter-pixel correlations, but are also influenced by stylistic noise from the training set, such as high-level semantics and color distribution. Among the previous works, this issue is overlooked even though necessary.

To address this issue, we introduce a novel deep superpixel method named the Content Disentangle Superpixel (CDS) algorithm to disentangle the dataset-specific style from the invariant content, i.e., the inter-pixel correlation used for superpixel segmentation. Our CDS achieve the objective of decoupling by constructing auxiliary data. As superpixel segmentation focuses not on high-level semantics but on pixel correlations, auxiliary data, despite altering style, do not disrupt the inherent pixel relationships. As illustrated in Fig.1(b), gradually modifying the stylistic information of both auxiliary and original data enhances the purity of the shared invariant inter-pixel correlations. Concretely, our method consists of three parts in order: feature extraction, content disentangle (CD), and superpixel generation. In the CD phrase, we feed the modality embeddings into the content selective gate to adaptively decouple content features and style noise. Afterwards, we propose the superpixel-grid correlation alignment to ensure that the selected content features from two modalities have the same pixel correlations. To prevent the occurrence of degenerate solutions in the content selective gate above. We enforce constraints to encourage the dataset/modal-specific features to have smaller mutual information while avoiding degradation of the selected modality style. Finally, a modality-shared superpixel decoder is designed to predict superpixel associations. Moreover, the auxiliary modality is only used in the training phase. In summary, our contributions are:

- We discover that existing deep superpixel algorithms depend on the distribution of training data and propose the CDS algorithm to ensure that learnable superpixels have a high generalization and boundary adherence power.
- During the training phase, CDS introduces an auxiliary modality to help decouple the correlated features among pure pixels in the RGB modality; while performing inference using only RGB. Compared to previous work, our algorithm does not increase additional computational burden but effective.
- Experimental results on datasets from four different domains indicate the superiority of our approach to existing superpixel algorithms. In addition, we demonstrate that our method improves the performance of downstream tasks.

Related Work

Traditional Superpixel Algorithms. The research of traditional superpixel algorithms (Achanta et al. 2012; Bergh et al. 2012; Felzenszwalb and Huttenlocher 2004; Li and Chen 2015; Liu et al. 2011; Grady 2006; Meyer 1992) has a long timeline. In times of scarce parallel computing resources, traditional superpixel algorithms can reduce image information redundancy and are often used to improve the computational efficiency of downstream tasks. Traditional superpixel algorithms are mainly classified into graph-based, clustering-based, and energy-based approaches. Concretely, graph-based methods, e.g., ERS (Liu et al. 2011), treat superpixel segmentation as a graph partitioning problem. The image is formulated as an undirected graph, and the edge weight indicates the inter-pixel similarity. Clustering-based methods, e.g., SLIC (Achanta et al. 2012), SNIC (Achanta and Susstrunk 2017), and LSC (Li and Chen 2015), initial the superpixels with seed pixels and apply cluster algorithms like k-means to adjust the pixel association. Energy-based methods, e.g., SEEDS (Bergh et al. 2012) and ETPS (Yao et al. 2015), first partition the image into regular grids and leverage different energy functions as an objective to exchange the pixels between neighboring superpixels. Traditional superpixel algorithms usually use CIELAB colors, concatenated with two-dimensional position encoding as pixel features.

Deep Superpixel Algorithms. SEAL (Tu et al. 2018) combines the neural networks with the ERS (Liu et al. 2011) by proposing the segmentation-aware affinity loss to learn cluster-friendly features. SEAL is not end-to-end trainable. To address this problem, Jampani *et al.* relaxes the nearest neighbor constraints of SLIC and develops the first differentiable deep algorithm SSN (Jampani et al. 2018). Since both SEAL and SSN require iterative traditional superpixel operations to complete the segmentation, SCN (Yang et al. 2020) model the superpixel segmentation as a classification problem between each pixel and its neighboring nine superpixels, which significantly improves the computational efficiency of superpixel segmentation. Based on the SCN, Wang *et al.* (Wang et al. 2021) propose the association implantation module, i.e., AINet, to enable the network to enhance the relations between the pixel and its surrounding grids. Among them, SCN and AINet are the SOTA algorithms with optimal performance.

Style Removal Algorithms. To better understand our work, we introduce the style removal techniques related to our Motivation. Style removal is not an independent visual task, often used as a technical means to solve problems such as style transfer and domain generalization. Existing methods mainly focus on two aspects: Normalization (Pan et al. 2018; Ulyanov, Vedaldi, and Lempitsky 2017; Huang and Belongie 2017) and Whitening (Li et al. 2017; Pan et al. 2019; Cho et al. 2019; Choi et al. 2021). Especially in (Ulyanov, Vedaldi, and Lempitsky 2017), the authors propose instance normalization to prevent overfitting on the domain-specific style of training data. (Pan et al. 2018) achieve significant performance improvement by incorporating the IN layers to capture style-invariant information. (Li et al. 2017) visually validates that the image style exists mainly on the correlation of feature channels and proposes the whitening transform to extract image content. Channel whitening calculates the covariance matrix across all channels, resulting in excessive computational complexity ($O(n^2)$). In (Cho et al. 2019), the authors introduce group-wise instance whitening (GIW)

transform to improve time efficiency. Overall, instance-level normalization and whitening both ensure that feature channels are independent of each other.

Preliminaries

Superpixel segmentation is the task of partitioning an input image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ into a set of n superpixels $\mathcal{S} = \{S_1, S_2, ..., S_n\}$, where each superpixel S_i consists of a group of adjacent pixels with similar characteristics. Mathematically, superpixel segmentation algorithms aim to calculate the pixel-superpixel association map $\mathcal{Q}$. Since computing the association between each pixel p with n superpixels has a high computational complexity, the latest deep superpixel approaches (Yang et al. 2020) formulate the superpixel segmentation as a local classification problem between pixel p with nine neighbor superpixel grids, and improve the time efficiency. This also served as our theoretical foundation.

Concretely, the image $\mathcal{I}$ is initialized into regular grids, and $\mathcal{Q} \in \mathbb{R}^{H\times W\times 9}$ indicates the probability that each pixel p is attributed to its nine nearby superpixel grids. The association map $\mathcal{Q}$ is predicted by a series of CNNs. Since no label is available for this output, deep algorithms design the superpixel loss inspired by the k-means convergence condition. Formally, let $f(p)$ be the pixel properties, i.e., semantic one-hot vector and positional encoding; we first obtain the superpixel clustering center properties $g(s)$ leveraging $\mathcal{Q}$ and $f(p)$. Then reconstruct the pixel property as follows:

$$g(s) = \frac{\sum_{\{p|s\in N_p\}} f(p)\cdot \mathcal{Q}_s(p)}{\sum_{\{p|s\in N_p\}} \mathcal{Q}_s(p)}, f'(p) = \sum_{s\in N_p} g(s)\cdot \mathcal{Q}_s(p). \tag{1}$$

Here N_p indicates the set of adjacent superpixels of p, and the $\mathcal{Q}_s(p)$ is the predicted probability that pixel p is assigned to superpixel s. Finally, superpixel segmentation amounts to minimizing the reconstruction distance

$$L_{sp}(Q) = \sum_p \mathrm{Dis}\left(f(p), f'(p)\right). \tag{2}$$

Following previous works (Yang et al. 2020; Wang et al. 2021), we use the cross-entropy and the Euclidean distance as the distance measure of the semantic label and the position vector, respectively.

Proposed Algorithm

In this section, we present Content Disentangle Superpixel (CDS) algorithm (shown in Fig.2), which aims to learn invariant inter-pixel correlations and reduce the style noise for superpixel generation. We first introduce the auxiliary modal and feature extraction, and then, present the content disentangle mechanism. Finally, the superpixel decoder is shown in the last subsection.

Auxiliary Modal and Feature Extraction

Auxiliary Modal. As shown in Fig.1(b), we construct samples with large domain offsets homogeneously sourced from the raw data to separate the inter-pixel correlation information, and the constructed auxiliary samples need to: *(1) preserve the pixel interrelations intact, (2) and exhibit significant stylistic differences.* Consequently, we adopt the auxiliary modality. Modality refers to the way in which something happens or is experienced (Baltrušaitis, Ahuja, and Morency 2018), and different modalities, i.e., HSV and LAB color-space transform and gradient map, describe objects from different perspectives. Although these modalities do not possess as significant modality barriers as heterogeneous data like text and images, they exhibit substantial variations in pixel-level descriptions.

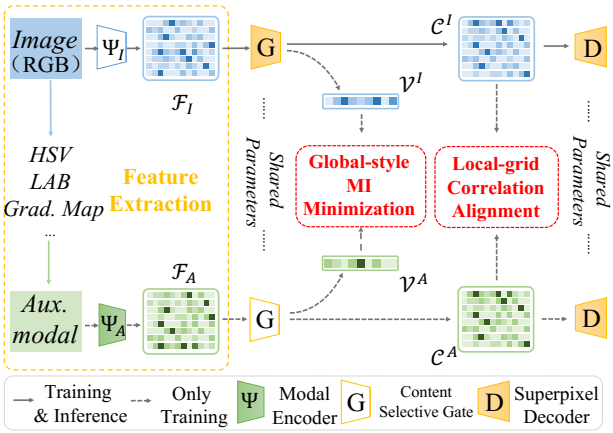


Figure 2: Flowchart of the proposed content disentangle superpixel algorithm.

Modal Encoder. Different from (Yang et al. 2020; Wang et al. 2021), to facilitate the decoupling of learnable image features, we remain in the explicit pixel-level feature extraction phase. For each image $\mathcal{I}$ and its auxiliary modality sample $\mathcal{A} \in \mathbb{R}^{H\times W\times 3}$, the whole process can be formulated as:

$$\Psi(\mathcal{I}) = \mathcal{P}\left\{\Phi\left(\mathcal{I};\theta\right) + \mathcal{I};\theta^*\right\} \tag{3}$$

$$\Psi(\mathcal{A}) = \mathcal{P}\left\{\Phi\left(\mathcal{A};\gamma\right) + \mathcal{A};\gamma^*\right\} \tag{4}$$

where the Φ is a series of convolution layers modified from the CNN part of SSN (Jampani et al. 2018). We adopt its structure for simplicity and efficiency. Since we formulate superpixels as a local classification problem, we do not concatenate positional encoding as SSN does. Additionally, to prevent the loss of pixel information, we additionally use a non-linear mapping function $\mathcal{P}\{\cdot\}$ to aggregate the original pixel values. θ, θ^* and γ, γ^* indicate the parameters of two modalities, respectively.

Content Disentangle Mechanism

After feature extraction process, we obtain the pixel-level embedding of the inputs, i.e., $\mathcal{F}_\mathcal{I}, \mathcal{F}_\mathcal{A} \in \mathbb{R}^{C\times H\times W}$. In this section, we introduce the content disentangle mechanism to adaptively select the inter-pixel correlation information, which contains three operations in detail.

Content Selective Gate is a learnable filter to separate the pixel embedding $\mathcal{F}_{\mathcal{A},\mathcal{I}}$ into content features $\mathcal{C}^{\mathcal{A},\mathcal{I}} \in \mathbb{R}^{C\times H\times W}$ and global modality/dataset style vectors $\mathcal{V}^{\mathcal{A},\mathcal{I}} \in$

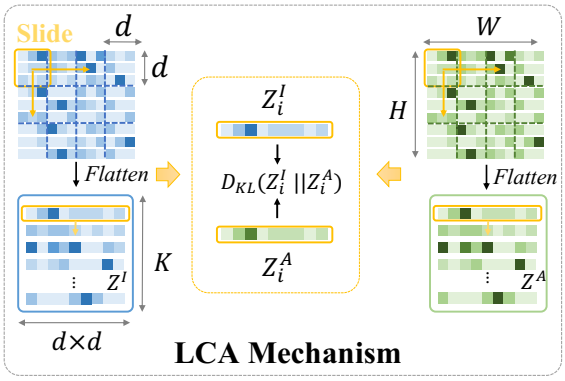


Figure 3: Illustration of the Local-grid Correlation Alignment (LCA) mechanism. LCA performs spatial domain distribution alignment at the superpixel level.

$\mathbb{R}^C$. As discussed in the related work, the image style exists mainly on the correlation of feature channels (Li et al. 2017). More importantly, superpixels do not consider global semantics like classification tasks do, and each position in the spatial domain has equal significance. Therefore, the content selective gate is designed as a share-weighted channel-wise attention strategy:

$$\begin{aligned} \mathcal{C}^i &= \mathrm{Gate}(\mathcal{F}_i) \cdot \mathcal{F}_i \\ \mathcal{V}^i &= \mathrm{avgpool}((1 - \mathrm{Gate}(\mathcal{F}_i)) \cdot \mathcal{F}_i) \end{aligned} \tag{5}$$

where $i \in \{\mathcal{A}, \mathcal{I}\}$, avgpool is the channel-wise average pooling operation, and

$$\mathrm{Gate}(\mathcal{F}_i) = \sigma\left(\mathrm{W}_2 \cdot \delta\left(\mathrm{W}_1 \cdot \mathrm{avgpool}(\mathcal{F}_i)\right)\right). \tag{6}$$

Here, W_1 and W_2 are the parameters for two fully-connected layers, and the $\delta(\cdot)$ and $\sigma(\cdot)$ are the relu and sigmoid function, respectively.

Local-grid Correlation Alignment is proposed to enforce the auxiliary modality and the primary modality to learn similar superpixel-friendly invariant features, i.e., the correlations between pixels. For superpixels, our objective is to ensure that embeddings of neighboring pixels with similar attributes are highly similar, while those of dissimilar pixels exhibit pronounced distinctions. This perspective highlights that superpixels prioritize capturing variations in feature dissimilarity at each image position, rather than being fixated on specific numerical values. Consequently, it becomes essential to guarantee the congruence of spatial distributions between the auxiliary and primary modalities.

As illustrated in Fig.3, given content features $\mathcal{C}^{\mathcal{I}}, \mathcal{C}^{\mathcal{A}} \in \mathbb{R}^{C \times H \times W}$ and initial superpixel grid with $d \times d$ size, We first employ average pooling to reduce dimensions, followed by traversing all superpixel grids and unfolding features in the spatial domain, resulting in the feature matrix $Z \in \mathbb{R}^{K \times d^2}$. K is the initial superpixel number which calculated by $(H/d) \times (W/d)$. Then, the local-grid correlation alignment constraint is formulated as:

$$\mathcal{L}_{align} = \frac{1}{K} \sum_{i}^{K} \mathcal{D}_{KL}(Z_i^I \| Z_i^A). \tag{7}$$

Algorithm 1: Training pseudocode for CDS.

Input: Input image $\mathcal{I}$ and auxiliary modal $\mathcal{A}$.
Output: The superpixel association map $\mathcal{Q}_{\mathcal{I}}$ and $\mathcal{Q}_{\mathcal{A}}$.
Initialize components: $\Psi_{\mathcal{I}}$, $\Psi_{\mathcal{A}}$, Gate, and superpixel decoder D. Initialize the variational distribution network h^θ.
for each training iteration **do**
  Step 1: Update the main network. (Fix the h^θ.)
  Conduct auxiliary modal $\mathcal{A}$.
    Calculate the $\mathcal{F}_{\mathcal{I}}$ and $\mathcal{F}_{\mathcal{A}}$ by Eq.3, Eq.4.
    Calculate the $\mathcal{C}^i$ and $\mathcal{V}^i$ by Eq.5, Eq.6.
    Predict the $\mathcal{Q}_{\mathcal{I}}$ and $\mathcal{Q}_{\mathcal{A}}$ by feeding $\mathcal{C}^i$ into the superpixel decoder $\mathcal{D}$.
    Calculate $\mathcal{L}_{align}$, $\mathcal{L}_{MI}$, and $\mathcal{L}_{sp}$, respectively.
    Update the CDS
  Step 2: Update the variational distribution network.
    Detach the value of $\mathcal{V}^I$ and $\mathcal{V}^A$ from the computational graph.
    Calculate $\mathcal{L}(\theta)$.
    Update the parameters of h^θ
end for

Here, we do not directly compute the global spatial distribution for three main reasons: (1) To enhance parallel computing efficiency. (2) To prevent the loss of local information, as softmax normalization is employed prior to calculating KL divergence, and a large number of pixels could lead to excessively small local probabilities. (3) The deep superpixel algorithm is defined as a neighborhood classification problem.

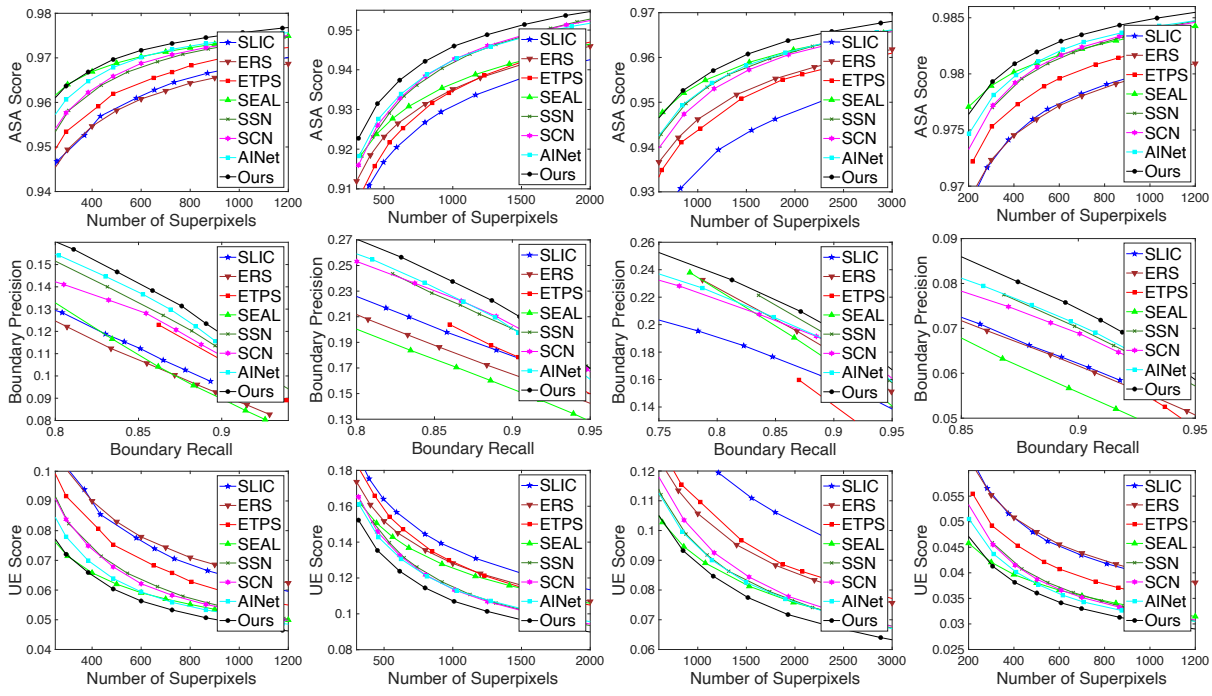
Global-style Mutual Information Minimization. Since only alignment lead to information loss or confusion between modalities, to prevent the occurrence of degenerate solutions in the content selective gate above, we enforce constraints to encourage the dataset/modal-specific features $\mathcal{V}$ to have smaller mutual information. Mathematically, given the style vectors $\mathcal{V}^{\mathcal{A},\mathcal{I}}$, (Cheng et al. 2020) provide an upper bound on their mutual information:

$$\begin{aligned} \mathrm{I}(\mathcal{V}^{\mathcal{I}}; \mathcal{V}^{\mathcal{A}}) \leq & \mathbb{E}_{p(\mathcal{V}^{\mathcal{A}}, \mathcal{V}^{\mathcal{I}})}[\log p(\mathcal{V}^{\mathcal{I}} \mid \mathcal{V}^{\mathcal{A}})] \\ & - \mathbb{E}_{p(\mathcal{V}^{\mathcal{A}})} \mathbb{E}_{p(\mathcal{V}^{\mathcal{I}})}[\log p(\mathcal{V}^{\mathcal{I}} \mid \mathcal{V}^{\mathcal{A}})] \end{aligned} \tag{8}$$

where $\mathrm{I}(\mathcal{V}^{\mathcal{I}}; \mathcal{V}^{\mathcal{A}})$ indicates the mutual information. However, the $\mathcal{V}^{\mathcal{A},\mathcal{I}}$ are learnable variables, and conditional distribution $p(\mathcal{V}^{\mathcal{I}} \mid \mathcal{V}^{\mathcal{A}})$ is unavailable. Consequently, we leverage the variational distribution h^θ to approximate $p(\cdot|\cdot)$. Then, the proposed MI loss is given to minimize the MI upper bound:

$$\mathcal{L}_{MI} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} [(\log h^\theta(\mathcal{V}_i^{\mathcal{I}} | \mathcal{V}_i^{\mathcal{A}})) - (\log h^\theta(\mathcal{V}_j^{\mathcal{I}} | \mathcal{V}_i^{\mathcal{A}}))] \tag{9}$$

Specifically, the variational distribution $h^\theta(\mathcal{V}^{\mathcal{I}}|\mathcal{V}^{\mathcal{A}}))$ is usually implemented with neural networks and optimized independently by $\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} h^\theta(\mathcal{V}_i^{\mathcal{I}} | \mathcal{V}_i^{\mathcal{A}})$, the log-likelihood function (Yue et al. 2022). When participating



(a) BSDS500 (b) NYUv2 (c) KITTI (d) VOC2012

Figure 4: Performance comparison on four datasets from different domains. From Left to Right: BSDS, NYU, KITTI and VOC datasets. From Top to Bottom: ASA, BR-BP and UE metrics. Except UE, higher values indicate that the algorithm is more effective.

in the computation of $\mathcal{L}_{MI}$ and updating of the main network, the parameters of h^θ are frozen, and only gradients are passed.

Superpixel Generation

Since the features encoding local pixel correlations, i.e., $\mathcal{C}^{\mathcal{I},\mathcal{A}}$ have the same size as the original image, we initially formulate hierarchical features $\{c_1, c_2, ..., c_n\}$, where $n = \log_2(d)$, employing bilinear interpolation. The size of c_i is $\frac{1}{2^{i-1}} \cdot (H, W), i \in [1, n]$. Then, similar to U-net, we construct the decoder in a bottom-up manner, and the predict association map $\mathcal{Q}$ is given as:

$$\mathcal{Q}_i = D(c_1, c_2, ..., c_n; \epsilon), \quad \forall i \in \{\mathcal{A}, \mathcal{I}\}, \tag{10}$$

where ϵ indicates the parameters of D. In this way, we extend the network's receptive field to encompass the adjacent vicinity of nine $d \times d$ superpixel grids, thereby enabling a more nuanced prediction of the relationship map.

$$\mathcal{L}_{total} = \mathcal{L}_{align} + \mathcal{L}_{MI} + \mathcal{L}_{sp}(\mathcal{Q}_\mathcal{I}) + \mathcal{L}_{sp}(\mathcal{Q}_\mathcal{A}) \tag{11}$$

Finally, the overall optimization objective of the proposed CDS is defined in Eq. 11, which comprises three components: the loss for the proposed local-grid correlation alignment, the mutual information loss, and the superpixel loss. Algorithm 1 demonstrates the pseudocode for training our algorithm.

Experiments

Experiment Settings

Dataset. Since the proposed CDS superpixel algorithm is introduced to mitigate the influence of attribute noise from the training set. We evaluate our method on four segmentation datasets from different domains: BSDS500 (Arbelaez et al. 2010), NYUv2 (Silberman et al. 2012), KITTI (Geiger, Lenz, and Urtasun 2012), Pascal VOC2012 (Everingham et al. 2015). Concretely, BSDS is an object edge detection dataset that contains five segmentation labels from different annotators for each image. NYUv2 is a semantic segmentation dataset for indoor scenes. KITTI is a commonly used street scene segmentation dataset for autonomous driving. Pascal VOC is a benchmark segmentation dataset with twenty object categories, which can be used for tasks such as object detection, instance segmentation.

Evaluation protocol. Follow the evaluation protocol of previous works (Yang et al. 2020; Wang et al. 2021), we only train our model on the BSDS500 dataset and run inference on the other datasets. The clustering performance is mainly evaluated by three public metrics: Achievable Segmentation Accuracy (ASA), Boundary Recall-Precision (BR-BP) curve, Under-segmentation Error (UE).

Implementation details. During the training phase, we apply data augmentation through random resize, random cropping to 208×208, and random horizontal/vertical flipping for our CDS. We trained the models using Adam optimizer.

Model	Time(ms)	Device
SLIC(Achanta et al. 2012)	120	CPU
ERS(Liu et al. 2011)	940	CPU
ETPS(Yao et al. 2015)	82	CPU
SEAL(Tu et al. 2018)	2658	GPU&CPU
SSN(Jampani et al. 2018)	278	GPU
SCN(Yang et al. 2020)	5	GPU
AINet(Wang et al. 2021)	29	GPU
Ours	6	GPU

Table 1: Runtime comparison for generating about 600 superpixels on NYUv2 with image size 608×448.

The learning rate starts at 5e-4 and is updated by the poly learning rate policy. We trained our model for 150k iterations with batch size eight, and superpixel grid size d is set to 16. We use the gradient map as the auxiliary modality and conduct all the experiments on single RTX3090 GPU.

Comparison with the State-of-the-Arts

Fig.4 and Tab.1 report the metrics and runtime comparisons with representative superpixel algorithms, respectively, including three traditional approaches, i.e., clustering-based method SLIC (Achanta et al. 2012), graph-based method ERS (Liu et al. 2011), and energy-based method ETPS (Yao et al. 2015), and four deep superpixel approaches, i.e., SEAL (Tu et al. 2018), SSN (Jampani et al. 2018), SCN (Yang et al. 2020) and AINet (Wang et al. 2021). For traditional superpixels, we leverage the hyperparameters posted by (Stutz, Hermans, and Leibe 2018). For deep superpixel algorithms, we conduct the experiment with their official implementations. Among them, SCN and AINet are the SOTA algorithms with optimal performance.

For a fair comparison, we employ the actual number of generated superpixels due to potential variations from manually specified counts. Our observations reveal the following:

- (1) Our method consistently outperforms others across all datasets, with a widening and stabilizing lead as superpixel count increases. This implies that style noise reduction enhances contour accuracy.
- (2) Across diverse domains, our superiority becomes more pronounced, highlighting superior generalization.
- (3) The auxiliary modality only participates in computations during the training phase, ensuring both improved algorithm performance and preserved inference speed.

Qualitative comparison. Fig.7 shows the qualitative results of six state-of-the art methods on dataset BSDS, NYUv2, KITTI, and Pascal VOC. Our method has better performance when facing critical object contours. Please see more visual results in the supplementary.

Ablation Study

We conduct experiments primarily to address the following key questions:

- **Q1**: The choices of auxiliary modalities, and the difference from using them as a form of data augmentation.

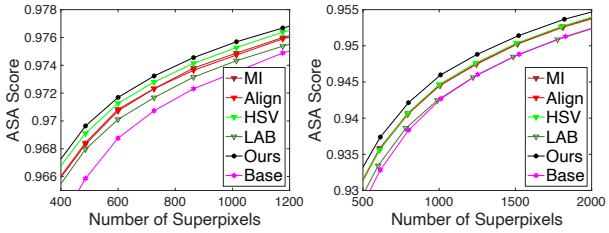


Figure 5: Component analysis. From left to right: ASA score on the BSDS and NYU datasets.

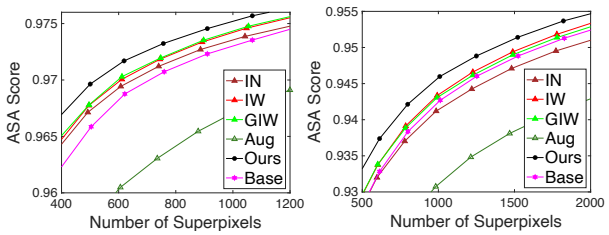


Figure 6: Comparison with other style removal methods. From left to right: ASA score on the BSDS and NYU datasets.

- **Q2**: Since our goal is to extract shared information from image pairs with significant style differences, why not use alignment directly.
- **Q3**: Can other universal style-removal methods improve the performance of the CDS model structure.

Component Analysis. As illustrated in Fig.5, we first study the choice of auxiliary modality, i.e., HSV, LAB, and gradient map (Ours), for question **Q1**. The performance ranking is Ours $>$ HSV $>$ LAB. As expected, the dissimilarity between the auxiliary modality and the RGB modality is positively correlated with the experimental performance. The LAB color space retains two color channels, while the image gradient map does not possess the same color information as RGB. Then, to answer the question **Q2**, the conducted experiments show that employing either feature alignment constraints or style mutual information minimization alone reduces the performance of the method, but both are still superior to the baseline method(Yang et al. 2020). We consider that (1) When only using alignment constraints, it actually does not decouple and remove the attribute noise of the dataset, but forcibly brings the two modalities closer in the feature space, resulting in the retention of some irrelevant information for superpixel segmentation. (2) Since the two modalities share the parameters of the superpixel generator, it is equivalent to implicit alignment. However, without the pixel correlation guidance for shared information extraction, it leads to a decrease in performance.

Comparison with universal style-removal approaches. We compare our method to other style-removal manners in Fig.6 to further verify the our effectiveness. Firstly, to supplement the question **Q1**, we employ the auxiliary modality transform as a data augmentation strategy (model Aug), i.e, applying the random modal transition by probability 0.3 for

(a) Image (b) Label (c) SLIC (d) SSN (e) SEAL (f) SCN (g) AINet (h) Ours

Figure 7: Visualization results of ours and previous methods. Compared to other superpixel algorithms, our method achieves better boundary adherence when facing unseen images. (From Top to Bottom: examples from BSDS, NYUv2, KITTI, VOC2012.)

training the single-modality model. Due to the significant style differences between modalities and the original RGB images, it confuses the lightweight feature extractor of the superpixel algorithm, resulting in a severe performance decline. Then we compare our methods with the IN (Ulyanov, Vedaldi, and Lempitsky 2017), IW (Li et al. 2017), GIW (Cho et al. 2019) for question **Q3**. As shown in the figure, our method achieve the best performance. Specifically, the generalization performance of Instance Normalization (IN) falls behind our baseline. We believe this is because during the training phase, Instance Normalization introduces additional errors due to significant individual variations.

Application

Superpixel algorithms provide better feature representation for downstream tasks, improving model performance and efficiency. The performance of superpixels in downstream tasks also demonstrates the effectiveness of superpixel algorithms. Table.2 reports the performance comparison on downstream semantic segmentation tasks. We leverage the pretrained Deeplab(Chen et al. 2014) and Bilateral Inception (BI) module(Gadde et al. 2016) and directly replace the superpixel segmentation generated by gSLICr(Ren, Prisacariu, and Reid 2015) with different superpixel results. The first row indicates the official implementation with 1000 superpixels. The following four lines show the results using 600 superpixels. Our method achieved an mIOU of 79.02 on the reduced validated set used by (Gadde et al. 2016) of Pascal VOC2012, outperforming other superpixel algorithms.

Base Modal	Methods	IoU
DeepLab	BI (gSLICr)	78.54
	BI (ETPS)	77.67
	BI (SCN)	78.90
	BI (AINet)	78.96
	BI (Ours)	79.02

Table 2: Superpixel algorithms with downstream segmentation. IoU comparison on the Pascal VOC dataset.

Conclusion

In this paper, we propose the Content Disentangle Superpixel algorithm to eliminate the dataset style noise that exists in the learnable superpixel features and reduce the feature-level distribution difference between training data and open-world data for superpixel segmentation. Unlike other deep superpixel algorithms that use single-modal training, we introduce auxiliary modalities to assist in decoupling the RGB image features into invariant image content and style noise. Specifically, we propose local-grid correlation alignment to maximize the selected image content information across modalities and learn the invariant inter-pixel correlations for superpixel generation. Then, we propose the global-style mutual information minimization to the minimize the upper bound of mutual information for style information, and prevent the degenerate solution during the disentangle process. Experimental results demonstrate that our method effectively mitigates the impact of style noise on deep superpixel algorithms and achieves superior results compared to existing methods on four different datasets.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (No.2021ZD0112100), and National Natural Science Foundation of China (No.61972022, No.U1936212, No.62120106009, No.52202486).

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11): 2274–2282.

Achanta, R.; and Susstrunk, S. 2017. Superpixels and polygons using simple non-iterative clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4651–4660.

Arbelaez, P.; Maire, M.; Fowlkes, C.; and Malik, J. 2010. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5): 898–916.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Bergh, M. V. d.; Boix, X.; Roig, G.; Capitani, B. d.; and Gool, L. V. 2012. Seeds: Superpixels extracted via energy-driven sampling. In *European conference on computer vision*, 13–26. Springer.

Birchfield, S.; and Tomasi, C. 1999. Multiway cut for stereo and motion with slanted surfaces. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 1, 489–495. IEEE.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, 1779–1788. PMLR.

Cho, W.; Choi, S.; Park, D. K.; Shin, I.; and Choo, J. 2019. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10639–10647.

Choi, S.; Jung, S.; Yun, H.; Kim, J. T.; Kim, S.; and Choo, J. 2021. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11580–11590.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136.

Felzenszwalb, P. F.; and Huttenlocher, D. P. 2004. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2): 167–181.

Gadde, R.; Jampani, V.; Kiefel, M.; Kappler, D.; and Gehler, P. V. 2016. Superpixel convolutional networks using bilateral inceptions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 597–613. Springer.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Grady, L. 2006. Random walks for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 28(11): 1768–1783.

Guo, Y.; Jiao, L.; Wang, S.; Wang, S.; Liu, F.; and Hua, W. 2018. Fuzzy superpixels for polarimetric SAR images classification. *IEEE Transactions on Fuzzy Systems*, 26(5): 2846–2860.

He, S.; Lau, R. W.; Liu, W.; Huang, Z.; and Yang, Q. 2015. Supercnn: A superpixelwise convolutional neural network for salient object detection. *International journal of computer vision*, 115: 330–344.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Jampani, V.; Sun, D.; Liu, M.-Y.; Yang, M.-H.; and Kautz, J. 2018. Superpixel sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 352–368.

Kwak, S.; Hong, S.; and Han, B. 2017. Weakly supervised semantic segmentation using superpixel pooling network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30.

Li, Z.; and Chen, J. 2015. Superpixel segmentation using linear spectral clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1356–1363.

Liu, M.-Y.; Tuzel, O.; Ramalingam, S.; and Chellappa, R. 2011. Entropy rate superpixel segmentation. In *CVPR 2011*, 2097–2104. IEEE.

Meyer, F. 1992. Color image segmentation. In *1992 international conference on image processing and its applications*, 303–306. IET.

Pan, X.; Luo, P.; Shi, J.; and Tang, X. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 464–479.

Pan, X.; Zhan, X.; Shi, J.; Tang, X.; and Luo, P. 2019. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1863–1871.

Ren, C. Y.; Prisacariu, V. A.; and Reid, I. D. 2015. gSLICr: SLIC superpixels at over 250Hz. *ArXiv e-prints*.

Sautier, C.; Puy, G.; Gidaris, S.; Boulch, A.; Bursuc, A.; and Marlet, R. 2022. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9891–9901.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, 746–760. Springer.

Stutz, D.; Hermans, A.; and Leibe, B. 2018. Superpixels: An evaluation of the state-of-the-art. *Computer Vision and Image Understanding*, 166: 1–27.

Tu, W.-C.; Liu, M.-Y.; Jampani, V.; Sun, D.; Chien, S.-Y.; Yang, M.-H.; and Kautz, J. 2018. Learning superpixels with segmentation-aware affinity loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 568–576.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6924–6932.

Wang, Y.; Wei, Y.; Qian, X.; Zhu, L.; and Yang, Y. 2021. AINet: Association Implantation for Superpixel Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7078–7087.

Wang, Z.-F.; and Zheng, Z.-G. 2008. A region based stereo matching algorithm using cooperative optimization. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.

Yang, F.; Sun, Q.; Jin, H.; and Zhou, Z. 2020. Superpixel segmentation with fully convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13964–13973.

Yao, J.; Boben, M.; Fidler, S.; and Urtasun, R. 2015. Real-time coarse-to-fine topologically preserving segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2947–2955.

Yue, L.; Liu, Q.; Du, Y.; An, Y.; Wang, L.; and Chen, E. 2022. DARE: Disentanglement-Augmented Rationale Extraction. *Advances in Neural Information Processing Systems*, 35: 26603–26617.

Zhao, C.; Zhu, W.; and Feng, S. 2022. Superpixel guided deformable convolution network for hyperspectral image classification. *IEEE Transactions on Image Processing*, 31: 3838–3851.

Zhu, W.; Liang, S.; Wei, Y.; and Sun, J. 2014. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2814–2821.