

А. О. Устюжанин¹, В. В. Китов²,
**СТИЛИЗАЦИЯ ИЗОБРАЖЕНИЙ С УЧЁТОМ
ПРОСТРАНСТВЕННОЙ ИНФОРМАЦИИ***

Постановка задачи

Задача переноса стиля (*style transfer*) заключается в автоматической перерисовке входного изображения, называемое *контентным изображением* или *контентом*, в стиле другого изображения, называемого *стилевым изображением* или *стилем*. Решение задачи позволяет, например, воспроизвести семейное фото в стиле картины известного художника. Многие онлайн-сервисы, такие как *ostagram.me* и *deepart.io*, а также мобильные приложения, такие как *Artisto*, *Prisma*, *Vinci*, используют эту технологию.

Результатом данной задачи является стилизованное изображение, сочетающее пространственную информацию с контентного изображения (изображенные на нём объекты должны сохранять узнаваемость) и стиль - со стилового изображения. Под стилем понимается распределение цветов, а также распределение локальных паттернов на изображении, таких как характерные мазки кисти.

Подобная задача решалась в методах, называемых *non-photorealistic photo rendering* [1, 2, 3], которые опирались на простые эвристики компьютерного зрения и не использовали нейросети. Однако их стилизующая способность была ограничена лишь некоторыми стилями, для которых заранее определялись соответствующие преобразования над изображением. Значительный вклад в развитие задачи переноса стиля внесла работа Леона Гатиса и др. [4], в которой использовались глубокие нейронные сети для воспроизведения стиля с произвольного стилового изображения. Преимуществом подхода является его гибкость работы с

¹разработчик программного обеспечения ООО «Яндекс.Технологии», e-mail: alex.ustyuzhanin98@yandex.ru.

²к.ф.-м.н., с.н.с. лаборатории искусственного интеллекта РЭУ имени Г. В. Плеханова, доцент факультета ВМК МГУ имени М. В. Ломоносова, e-mail: v.v.kitov@yandex.ru.

*Работа выполнена при финансовой поддержке государственного задания в сфере научной деятельности Министерства науки и высшего образования РФ на тему "Модели, методы и алгоритмы искусственного интеллекта в задачах экономики для анализа и стилизации многомерных данных, прогнозирования временных рядов и проектирования рекомендательных систем". Номер проекта FSSW-2023-0004.

произвольным стилем и силой стилизации, а недостатком — высокая вычислительная сложность алгоритма, требующая несколько десятков секунд на одну стилизацию даже на современных видеокартах.



Рис. 1. Пример работы задачи переноса стиля.

Ряд последующих методов производят стилизацию посредством пропуска контентного изображения через предобученную нейронную сеть, именуемую *трансформерной сетью* [5, 7, 6, 8, 9, 10]. Стилизация изображения требует единственного прохода через сеть и позволяет производить стилизацию в реальном времени, что, однако, отражается на качестве результата.

Общий недостаток стандартных подходов заключается в равномерном наложении стиля на изображение, игнорируя семантику контентного изображения. На практике же художники придают индивидуальный стиль и акцент разным объектам на картине: людям, сооружениям, ландшафту, небу и так далее. Проведенные в работе эксперименты демонстрируют, что четкое разделение основных объектов картины при стилизации играет ключевую роль в повышении художественной выразительности.

Работы [11, 7, 8] предлагают модификации, способные учитывать пространственную информацию в процессе переноса стиля. Недостатком этих подходов является необходимость вручную задавать маски объектов контентного изображения и маски подстилей, на которые разбивается требуемый стиль, а также вручную задавать сопоставление масок подстилей маскам объектов стилизуемого изображения. В этом смысле процедуры стилизации перестают быть автоматическими.

Алгоритмы стилизации, которые автоматически производят указанные шаги называют методами *мультимодальной стилизации* (*multimodal style transfer*). В данной работе предлагается обзор основных методов классической равномерной и мультимодальной стилизации, а также предлагается улучшенный алгоритм мультимодальной стилизации изображений, учитывающий пространственную информацию,

посредством автоматического извлечения сегментационных масок контентного изображения, и использующий для стилизации одно единственное стилевое изображение. Предлагаемый метод использует в качестве основы процедуру *WCT*, описанную в работе Ванга и др. [10]. В предложенном алгоритме также проводится кластеризация стилевого изображения, поскольку процедура нейростилевой сегментации, как правило, не предусматривает возможности работы с абстракциями, встречающимися в картинах художников. Выделенные сегменты контента и кластеры стиля предлагается сопоставлять между собой при помощи специальной процедуры, автоматически определяющей наилучшее соответствие. Итого, данный метод принимает на традиционную пару (контентное изображение, стилевое изображение), как и классические подходы стилизации, но накладывает стиль неравномерно по объектам изображений, чем производит более выразительные стилизации, как это можно увидеть на рисунке 2.



Рис. 2. Сравнение стилизаций предложенного метода с базовым *WCT*.

Постановка задачи

Задача переноса стиля принимает на вход стилевое изображение x_{style} , задающее интересующий нас художественный стиль, например, картину художника, и контентное изображение $x_{content}$, подлежащее перерисовке в требуемом стиле.

Выходное изображение x_o получается в результате оптимизации двух функций потерь: $\mathcal{L}^{content}$, называемой *контентной функцией потерь*, и *стилевой функцией потерь* $\mathcal{L}^{style}$. Контентная функция потерь нацелена на минимизацию различий между оптимизируемым изображением x и контентным изображением $x_{content}$, тем самым сохраняя исходное информационное наполнение. Стилeвая функция потерь позволяет перенести цветовую гамму и текстуру стилeвого изображения. Таким образом, оптимизационный процесс имеет вид:

$$x_o = \arg \min_x [\mathcal{L}^{content}(x_{content}, x) + \alpha \mathcal{L}^{style}(x_{style}, x)], \quad (1)$$

где α - неотрицательный гиперпараметр, отвечающий за интенсивность переноса стиля. Чем ниже этот показатель, тем выше схожесть стилизованного изображения и контентного.

Концептуально алгоритмы нейросетевой стилизации делятся на два класса: класс *offline* методов, которые для каждой пары $(x_{content}, x_{style})$ итеративно проводят оптимизационный процесс в пространстве изображений, результатом которого является изображение-стилизация x_o , и класс *online* методов, в которых обучается трансформерная нейронная сеть, продуцирующая стилизацию x_o стилевым изображением x_{style} за один проход по сети для произвольного изображения $x_{content}$. Между двумя данными классами возникает компромисс между качеством стилизации, которое, как правило, выше у *offline* методов, и скоростью получения стилизации, которая на порядок выше у *online* методов. Обратим внимание, что у *offline* методов оптимизация производится по пикселям оптимизируемой стилизации, в то время как у *offline* методов оптимизация ведется по весам трансформерной сети.

Обзор литературы

Фундаментальную роль в области задачи переноса стиля играет работа Гатиса и др. [4]. Данное исследование основывается на использовании сверточных нейронных сетей. Указанная сеть принимает на вход некоторое изображение, пропускает его через серию промежуточных сверточных слоев (фильтров) предобученной сверточной сети VGG [12] и возвращает извлеченные признаки информационного наполнения изображения. Стилизованное изображение получается в результате минимизации взвешенной суммы контентной и стилевой функций потерь:

$$\mathcal{L}^{content}(x_{content}, x) + \alpha \mathcal{L}^{style}(x_{style}, x) \longrightarrow \min_x, \quad (2)$$

$$\mathcal{L}^{content}(x_{content}, x) = \frac{1}{W_{l_c} H_{l_c} C_{l_c}} \|F_{l_c}(x_{content}) - F_{l_c}(x)\|_F^2, \quad (3)$$

$$\mathcal{L}^{style}(x_{style}, x) = \sum_{l \in l_s} \frac{1}{C_l^2} \|G_l(x_{style}) - G_l(x)\|_F^2, \quad G_l(x) = \frac{1}{W_l H_l} F_l(x) F_l^\top(x), \quad (4)$$

где сверточные представления $F_l(x) \in \mathbb{R}^{C_l \times W_l \times H_l}$ извлечены из промежуточного представления со слоя l нейронной сети, как правило VGG-16 [12], W_l и H_l - пространственные размерности признаков, C_l - число фильтров. Предлагается использовать множества слоев нейронной сети $l_c = \{\text{'relu2_2'}\}$ для контентной функции потерь и набор слоёв $l_s = \{\text{'relu1_1'}, \text{'relu2_1'}, \text{'relu3_1'}, \text{'relu4_1'}\}$ для стилевой.

Контентная функция потерь отвечает за сохранение характеристик объектов, представленных на исходном контентном изображении. Стилловая функция потерь штрафует художественные отличия между

оптимизируемой стилизацией и стиливым изображением. Это осуществляется при помощи вычисления нормы разности соответствующих матриц Грама, отвечающих извлеченным сверточным признакам стилизации и стиля. Минимизация обоих функционалов соответствует разумному балансу между сохранением содержимого контентного изображения и переносом художественных признаков стиливого изображения. Выходное изображение получено в результате итеративной оптимизации в пространстве пикселей изображения, которая требует несколько сотен итераций с высокой вычислительной сложностью. Подобная процедура занимает несколько десятков секунд даже на современных видеокартах. Ввиду вычислительной сложности данного подхода, в дальнейшем применяется иная процедура *online* стилизации, заключающаяся в обучении трансформерной сети, способной в реальном времени стилизовать произвольные контентные изображения некоторым единственным стилем, либо ограниченным множеством стилей.

WCT стилизация. Идея *WCT* стилизации [10] заключается в кодировании входного изображения при помощи предобученной нейронной сети (первые слои *VGG-19* [12]) в его сверточное представление, применении *whitening-coloring* преобразования, стилизующего внутреннее представление входного изображения в требуемом стиле, затем декодировании обратно в пиксельное пространство посредством предобученного декодировщика, имеющего зеркальную структуру по отношению к кодировщику.

Обозначим $f_c \in \mathbb{R}^{C \times H_c \times W_c}$ and $f_s \in \mathbb{R}^{C \times H_s \times W_s}$ за сверточные признаки сети *VGG-19* на некотором слое, например ‘relu4_1’, контентного изображения x_c и стиливого изображения x_s соответственно.

Вначале применяется *whitening* трансформация, которая осуществляет удаление исходного стиля для внутреннего представления входного контентного изображения за счёт нормализации:

$$\hat{f}_c = E_c D_c^{-\frac{1}{2}} E_c^\top \tilde{f}_c, \quad (5)$$

где $\tilde{f}_c = f_c - m_c$ — центрированные сверточные признаки контента, полученного с некоторого слоя предобученной сети *VGG*, D_c — диагональная матрица собственных значений матрицы ковариации $\frac{1}{H_c W_c} \tilde{f}_c \tilde{f}_c^\top \in \mathbb{R}^{C \times C}$, а E_c — соответствующая ей ортогональная матрица собственных векторов, удовлетворяющая равенству $\frac{1}{H_c W_c} \tilde{f}_c \tilde{f}_c^\top = E_c D_c E_c^\top$.

Преобразованные признаки $\hat{f}_c$ имеют нулевое математическое ожидание и единичную матрицу ковариации ($\frac{1}{H_c W_c} \hat{f}_c \hat{f}_c^\top = I$).

Затем применяется *coloring* трансформация, которая накладывает требуемый стиль на внутреннее представление контента за счёт смещения и масштабирования. Она применяет линейное преобразование к

центрированным и декоррелированным признакам контента $\hat{f}_c$, результат которого имеет математическое ожидание и матрицу ковариации, совпадающими с соответствующими статистиками признаков стилового изображения:

$$\hat{f}_{cs} = E_s D_s^{\frac{1}{2}} E_s^\top \hat{f}_c + m_s, \quad (6)$$

где D_s - диагональная матрица собственных значений матрицы ковариации $\frac{1}{H_s W_s} \tilde{f}_s \tilde{f}_s^\top \in \mathbb{R}^{C \times C}$, где $\tilde{f}_s = f_s - m_s$ - центрированные сверточные признаки стиля, а E_s - соответствующая ей ортогональная матрица собственных векторов.

В результате стилизованные признаки $\hat{f}_{cs}$ имеют ту же матрицу ковариации и среднее, что и стиливые признаки, за счёт чего достигается стилизация.

Multimodal Style Transfer via Graph Cuts (MST) *Multimodal style transfer via graph cuts (MST)* подход [13] принадлежит к классу мультимодальных методов стилизации. В нём формулируется задача установления соответствия между кластерами стиля и признаками контента в сверточном пространстве признаков изображения как задача минимизации функционала энергии на графе, решаемая при помощи поиска оптимального разреза графа (*graph cuts*). Авторы предлагают кластеризовать стиливое изображение и адаптировать процедуру стилизации под пространственную конфигурацию контента.

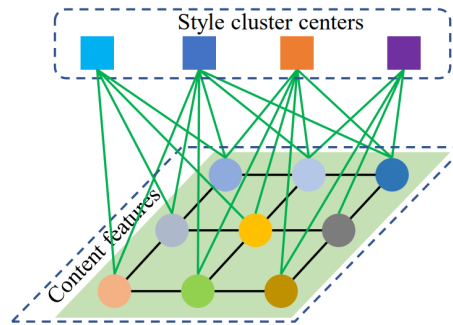


Рис. 3. Процедура сопоставления контент-стиль в алгоритме *MST*.

Данный алгоритм имеет следующую последовательность действий:

1. Извлекаются сверточные признаки контента $F_c \in \mathbb{R}^{C \times H_c W_c}$ и стиля $F_s \in \mathbb{R}^{C \times H_s W_s}$ при помощи кодировщика $E_{\theta_{enc}}(\cdot)$ (например, *VGG*).
2. Применяется алгоритм *K-средних* для кластеризации стиливых признаков F_s в K кластеров:

$$F_s = F_s^{l_1} \cup F_s^{l_2} \cup \dots \cup F_s^{l_K} \quad (7)$$

3. Вычисляется косинусное расстояние между пространственными точками признаков контента $F_{c,p} \in \mathbb{R}^{C \times 1}$ и центрами кластеров стиля

$$F_{s,l_k} \in \mathbb{R}^{C \times 1}, \quad l_k \in \{l_1, l_2, \dots, l_K\}:$$

$$D(F_{c,p}, F_{s,l_k}) = 1 - \frac{F_{c,p}^\top F_{s,l_k}}{\|F_{c,p}\| \|F_{s,l_k}\|} \quad (8)$$

4. Устанавливается каждой пространственной точке $F_{c,p}$ соответствие с номером стилового кластера при помощи функции сопоставления $f(p) \rightarrow \{l_1, l_2, \dots, l_K\}$. Функция f вычисляется в результате решения оптимизационной задачи *об оптимальном разрезе графа (graph cuts)*.

5. Группируются пространственные точки признаков контента по соответствию одному стиловому кластеру:

$$F_c = F_c^{l_1} \cup F_c^{l_2} \cup \dots \cup F_c^{l_K} \quad (9)$$

6. Преобразования *whitening* и *coloring* проводятся для каждой пары $\{F_c^{l_k}, F_s^{l_k}\}$ для получения признаков стилизации $F_{cs}^{l_k}$ кластером l_k :

$$F_{cs} = F_{cs}^{l_1} \cup F_{cs}^{l_2} \cup \dots \cup F_{cs}^{l_K} \quad (10)$$

7. Полученные признаки стилизации F_{cs} подаются в декодировщик $D_{\theta_{dec}}(\cdot)$ для реконструкции итогового стилизованного изображения I_{cs} .

Таким образом, данный алгоритм производит стилизацию, в которой ярко и четко выражены индивидуально стилизованные сегменты, что способствует повышению качества восприятия картины.

Structure-Emphasized Multimodal Style Transfer (SEMST)

Structure-emphasized multimodal style transfer (SEMST) подход [14] идейно наиболее близок к предлагаемому в данной работе алгоритму. В исследовании [14] авторы раскрывают некоторые проблемы, возникающие у ранее упомянутого подхода *MST* [13], и предлагают следующие модификации:

- Использование алгоритма кластеризации *K*-средних вместе с простым итеративным алгоритмом автоматического определения числа кластеров *K*, что значительно упрощает процесс настройки модели под разные контенты и стили.
- Кластеризация и стилового, и контентного изображений, что увеличивает гибкость стилизации, а также обеспечивает непрерывность построенных соответствий между фрагментами изображений.
- Вместо промежуточных сверточных представлений *VGG* кластеризуются исходные пиксели изображений в цветовом пространстве *CIE LAB*, что повышает точность определения границ кластеров.

- Простая процедура сопоставления вместо оптимизации graph cuts. Сопоставление подразумевает соответствие $f_i(S) : S_{i_1}, S_{i_2}, \dots, S_{i_{M_i}} \rightarrow C_i, i = 1, |C|$ нескольких стиливых кластеров некоторому одному кластеру контента, причем один стиливой кластер может участвовать в стилизации многих контентных кластеров.

Процедура автоматического подбора числа кластеров для алгоритма К-средних имеет следующую идею: происходит разделение на кластера, пока наибольший из них превосходит по размеру заданную пороговую долю от общего размера изображения, размер определяется числом пикселей. Таким образом, алгоритм стартует с числа кластеров $K = 2$ и увеличивает это значение $K := K + 1$, пока все кластера не будут ограничены в размере некоторым параметром. Заметим, что указанный подход, в отличие от предлагаемого, не поддерживает случай единственного кластера.

Процедура сопоставления кластеров устроена естественным образом: рассчитываются центры кластеров в цветовом пространстве *CIELAB* для контента и стиля, затем между ними устанавливается соответствие, если абсолютная разница между выбранными центрами не превышает порог, определяемый гиперпараметром. Если некоторый кластер контента был сопоставлен нескольким кластерам стиля, то в этом случае они объединяются и участвуют в стилизации выбранной области с равными весами. Далее применяется процедура *WCT* к каждой сопоставленной паре.

Предлагаемый метод улучшенной мультимодальной стилизации

В данной работе предлагается новый алгоритм мультимодальной стилизации, который мы назовём *Improved Multimodal Style Transfer* или *IMST*, с разбиением контентного и стиливого изображения на блоки. Он построен на основе *WCT* стилизации [10], которая равномерно распределяет стиль по изображению без учета пространственной информации. В предлагаемом алгоритме, наоборот, контент и стиль разделяются на области, затем стилизация проводится направленным образом область-в-область, контролируемая специальной процедурой сопоставления. Код предложенного алгоритма выложен в свободный доступ¹.

¹<https://github.com/alexust98/Improved-Multimodal-Style-Transfer>

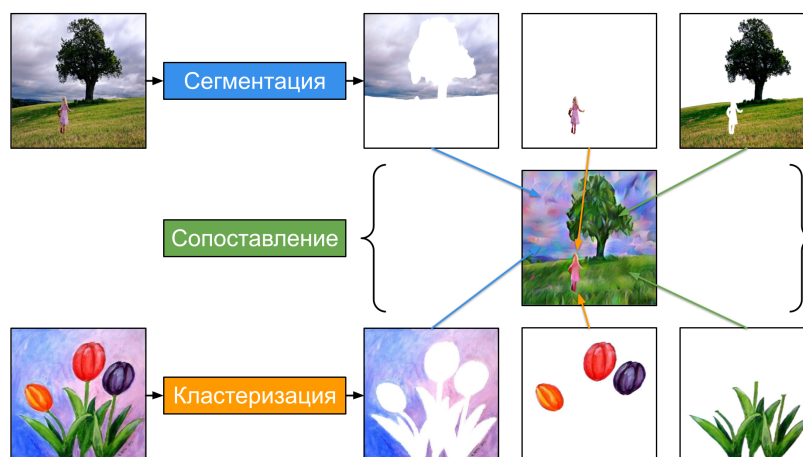


Рис. 4. Общая схема предлагаемого метода улучшенной мультимодальной стилизации.

Несмотря на то, что модифицированная мультимодальная стилизация алгоритмически построена поверх *WCT* стилизации [13], возможно аналогичным образом реализовать предложенные идеи с использованием произвольной трансформерной сети. Метод улучшенной мультимодальной стилизации предлагает следующие модификации и решения ранее выявленных проблем:

- Предлагается сегментация контентного изображения вместо кластеризации алгоритмом К-средних. Данная естественная модификация значительно увеличивает качество стилизации, благодаря высокой точности нейронных сегментаторов и факту разбиения контента на семантические сегменты, а не близкие по цвету области.
- Используется более гибкая процедура кластеризации, которая, в отличие от К-средних, позволяет автоматически определять необходимое число кластеров, а сами кластера могут быть невыпуклой формы.
- Предлагается устанавливать соответствие между сегментами и кластерами на основе матриц Грама, а не векторов средних. Исходя из проведенных экспериментов, выявлено, что это лучше описывает визуальную похожесть областей.
- Каждому сегменту контента сопоставляется смесь кластеров стиля, то есть каждый подстиль участвует в стилизации рассматриваемого сегмента с некоторым весом, в отличие от ранее рассматриваемого выбора равнозначных кластеров. Это добавляет гибкость в процедуру сопоставления областей.
- Вводится функция штрафов в процесс сопоставления областей, которая контролирует полноту переноса стиля. После вычисления

соответствия между сегментом контента и взвешенной суммой кластеров стиля, подстиль, имеющий, наибольший вес в стилизации штрафуются, чтобы избежать стилизации одним “хорошим” кластером и добиться более полного учёта информации со всего стилового изображения.

Для того, чтобы художественно акцентировать внимание на отдельных объектах контентного изображения предлагается проводить его семантическую сегментацию вместо кластеризации на непересекающиеся сегменты: $I_c = I_c^1 \cup I_c^2 \cup \dots \cup I_c^N$. Кластеризация позволяет сгруппировать визуально схожие области изображения, в то время как сегментация выделяет объекты на основе их семантической принадлежности одному классу, например, классу ‘человек’, ‘животное’, ‘транспорт’ и т.д.

Стилевое изображение разделяется на области при помощи кластеризации, поскольку процедура сегментирования не способна выделить абстрактные классы, часто встречающиеся в картинах художников: $I_s = I_s^1 \cup I_s^2 \cup \dots \cup I_s^M$. После этого мы извлекаем закодированные представления выделенных сегментов контента F_c^i и кластеров стиля F_s^j и применяем *WCT* процедуру для индивидуальной стилизации каждой сопоставленной пары.

Предлагаемый метод *IMST* формализован в алгоритме 1. Он ссылается на описанные далее процедуры *сегментации контента*, *кластеризации стиля* и *сопоставления областей*, необходимые для проведения *WCT* трансформаций применительно к выделенным парам сегментов-кластеров.

В отличие от методов *MST* и *SEMST*, в предложенном методе не сопоставляется подмножество кластеров стиля каждому сегменту контента, а рассчитываются нормированные веса каждому кластеру для улучшенного контроля над процессом стилизации. Эти веса учитываются в *WCT* преобразованиях, а именно при вычислении векторов средних и матриц ковариации, используемых на этапе *coloring*.

Сегментация контентного изображения. В целях сегментации используется композиция двух предобученных нейросетевых моделей: модель сегментации *DeepLabv3* [16] и модель оценки глубины для каждого пикселя изображения *MiDaS* [17].

Модель *DeepLabv3* обучена на выделение наиболее часто встречающихся объектов на изображении: людей, животных, транспорта и т.д. Модель оценки глубины *MiDaS* используется для разделения видимой части неба и земной поверхности на изображении. Области контента, которые не относятся ни к одному из классов модели *DeepLabv3* и имеют глубину выше заданного порога, определяются как

отдельный класс ‘небо’. Выбор порога не вызывает трудностей в силу резкого и значительного перепада значений оценки глубины изображения.

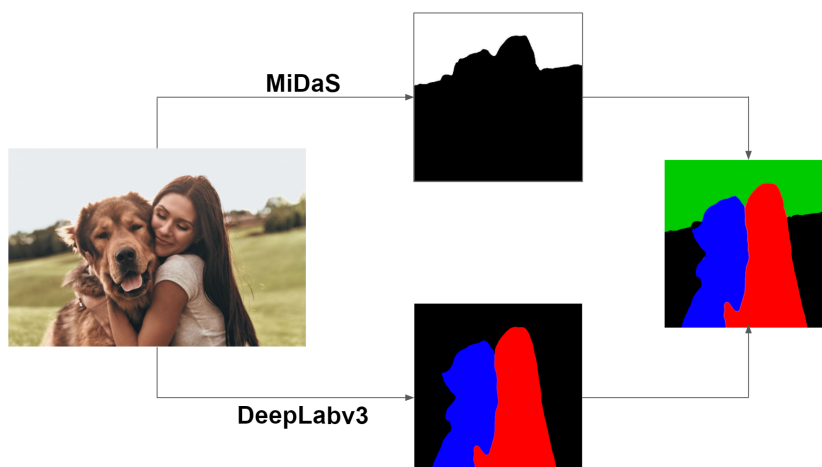


Рис. 5. Процедура сегментации контентного изображения.

Кластеризация стилового изображения. Для кластеризации стилизованных изображений используется алгоритм *HDBSCAN* [15], устойчивый к выбросам, автоматически определяющий число кластеров и имеющий возможность выделять невыпуклые области произвольной формы.

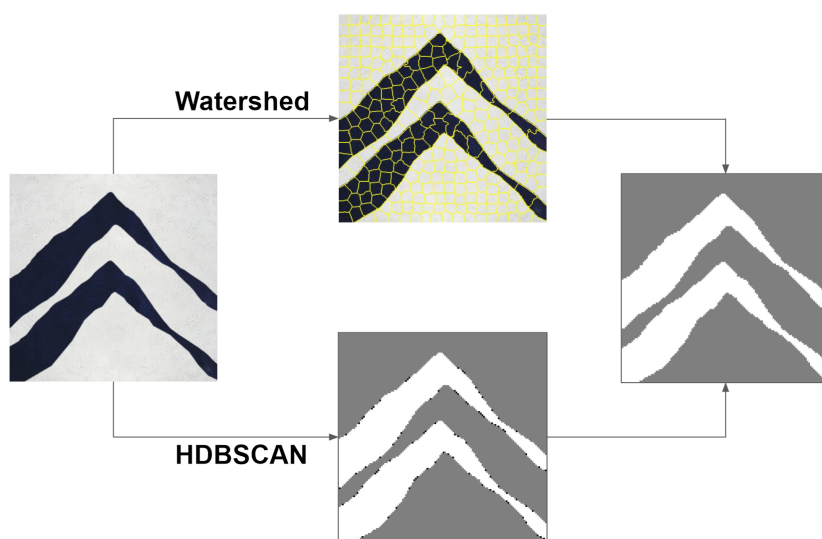


Рис. 6. Процедура кластеризации стилового изображения.

Для уточнения результата работы алгоритма кластеризации используется процедура постобработки: изображение покрывается неравномерной сеткой, разделяющей его на *суперпиксели* (связные области со схожим значением пикселей) при помощи алгоритма сегментации *Watershed* [18]. Каждый суперпиксель относится к одному из кластеров *HDBSCAN*, определяемому большинством своих пикселей. Таким образом, данная процедура уточняет границы кластеров, которые

могли быть отнесены к шумовому кластеру *HDBSCAN*, как представлено на рисунке 6.

Процедура сопоставления областей. Этап сопоставления сегментов контентного изображения и кластеров стилового изображения необходим для автоматического определения областей стиля, которые наилучшим образом стилизуют заданную область контента и показан в алгоритме 1.

Алгоритм 1 Предлагаемый метод улучшенной мультимодальной стилизации

Вход:

I_c - Изображение контента, I_s - Изображение стиля

$E_{\theta_{enc}}(\cdot)$ - Предобученный кодировщик (VGG-19)

$D_{\theta_{dec}}(\cdot)$ - Преобученный декодировщик (взятый из метода MST)

Выход:

I_{cs} - Стилизованное изображение

- 1: $F_c^{2,1}, F_c^{4,1} \leftarrow E_{\theta_{enc}}(I_c)['relu2_1', 'relu4_1']$ // Кодлируем контент
 - 2: $F_s^{2,1}, F_s^{4,1} \leftarrow E_{\theta_{enc}}(I_s)['relu2_1', 'relu4_1']$ // Кодлируем стиль
 - 3: $C \leftarrow \text{Сегментация}(I_c)$ // Сегментируем контент (DeepLabv3 + MiDaS)
 - 4: $S \leftarrow \text{Кластеризация}(I_s)$ // Кластеризуем стиль (HDBSCAN + Watershed)
 - 5: $W \leftarrow \text{Сопоставление}(C, S, F_s^{2,1}, F_c^{2,1})$ // Вычисляем соответствие областей контент-стиль
 - 6: $F_{cs}^{4,1} \leftarrow F_c^{4,1}$
 - 7: **Цикл** $i = 1, 2, \dots, |C|$:
 - 8: $f_{c_i} \leftarrow F_c^{4,1}[C_i] \in \mathbb{R}^{C_4 \times Q_i^{4,1}}$ // $Q_i^{4,1}$ - мощность контентного сегмента C_i на слое 'relu4_1'
 - 9: $\Sigma_{c_i} \leftarrow \frac{1}{Q_i^{4,1} - 1} (f_{c_i} - \bar{f}_{c_i})(f_{c_i} - \bar{f}_{c_i})^T \in \mathbb{R}^{C_4 \times C_4}$ // Матрица ковариации
 - 10: $U_{c_i}, \Lambda_{c_i}, V_{c_i} \leftarrow \text{SVD}(\Sigma_{c_i})$
 - 11: $\hat{f}_{c_i} \leftarrow V_{c_i} \Lambda_{c_i}^{-\frac{1}{2}} U_{c_i}^T (f_{c_i} - \bar{f}_{c_i})$ // Whitening трансформация
 - 12: $\mu_s \leftarrow \frac{\sum_{j=1}^{|S|} (F_s^{4,1}[S_j] * W_{ij})}{\sum_{j=1}^{|S|} (P_j^{4,1} * W_{ij})} \in \mathbb{R}^{C_4}$ // $P_j^{4,1}$ - размер стилового кластера S_j на слое 'relu4_1'
 - 13: $\Sigma_s \leftarrow \frac{\sum_{j=1}^{|S|} ((F_s^{4,1}[S_j] - \mu_s)(F_s^{4,1}[S_j] - \mu_s)^T * W_{ij})}{\sum_{j=1}^{|S|} (P_j^{4,1} * W_{ij}) - 1} \in \mathbb{R}^{C_4 \times C_4}$
 - 14: $U_s, \Lambda_s, V_s \leftarrow \text{SVD}(\Sigma_s)$
 - 15: $\hat{f}_{cs_i} \leftarrow U_s \Lambda_s^{\frac{1}{2}} V_s^T \hat{f}_{c_i}$ // Coloring трансформация
 - 16: $F_{cs}^{4,1}[C_i] = \hat{f}_{cs_i} + \mu_s$
 - 17: $I_{cs} \leftarrow D_{\theta_{dec}}(F_{cs}^{4,1})$
 - 18: **Вернуть** I_{cs}
-

Указанная процедура осуществляется посредством автоматического расчета весового коэффициента для каждого из стиливых кластеров, определяющего интенсивность соответствующего кластера при стилизации.

Обозначим множество сегментов контента как $C = \{C_1, C_2, \dots, C_N\}$ и множество стиливых кластеров как $S = \{S_1, S_2, \dots, S_M\}$. Предлагается следующая функция:

$$W(C_i, S) \rightarrow \{\omega_1, \omega_2, \dots, \omega_M\}, \quad i = 1, \dots, N, \quad (11)$$

которая сопоставляет каждому стиливому кластеру вещественное число ω_j , $\sum_{j=1}^M \omega_j = 1$. Взвешивание определяется метрикой схожести стиливого кластера по отношению к рассматриваемому сегменту контента в пространстве признаков сверточной нейронной сети.

Целью процедуры сопоставления является определение весовых коэффициентов w_j стиливых кластеров таким образом, что максимальные значения весов достигаются на наиболее подходящих выбранному сегменту контента подстилю при условии сохранения их визуальной различимости. Для этого вводится *функция штрафов*. После вычисления весовых коэффициентов для текущего сегмента предлагается добавлять штраф стиливому кластеру с наибольшим весом, тем самым заставляя алгоритм выбирать этот кластер с меньшим весом для следующего сегмента контента в целях повышения их различимости между собой. Веса кластеров определяются их метрикой расстояния до выбранного сегмента, поэтому функция штрафов прибавляет положительное число к расстояниям кластера, имеющего наименьший показатель метрики на текущей итерации, до всех сегментов контента. Порядок в котором производится обход сегментов влияет на результат. Предлагается начинать с сегмента контента с минимальным значением расстояния до какого либо кластера стиля, а потом брать сегменты в порядке ее возрастания. Этим мы обеспечиваем сохранение наилучшего сочетания контента и стиля.

Сравнительный анализ методов

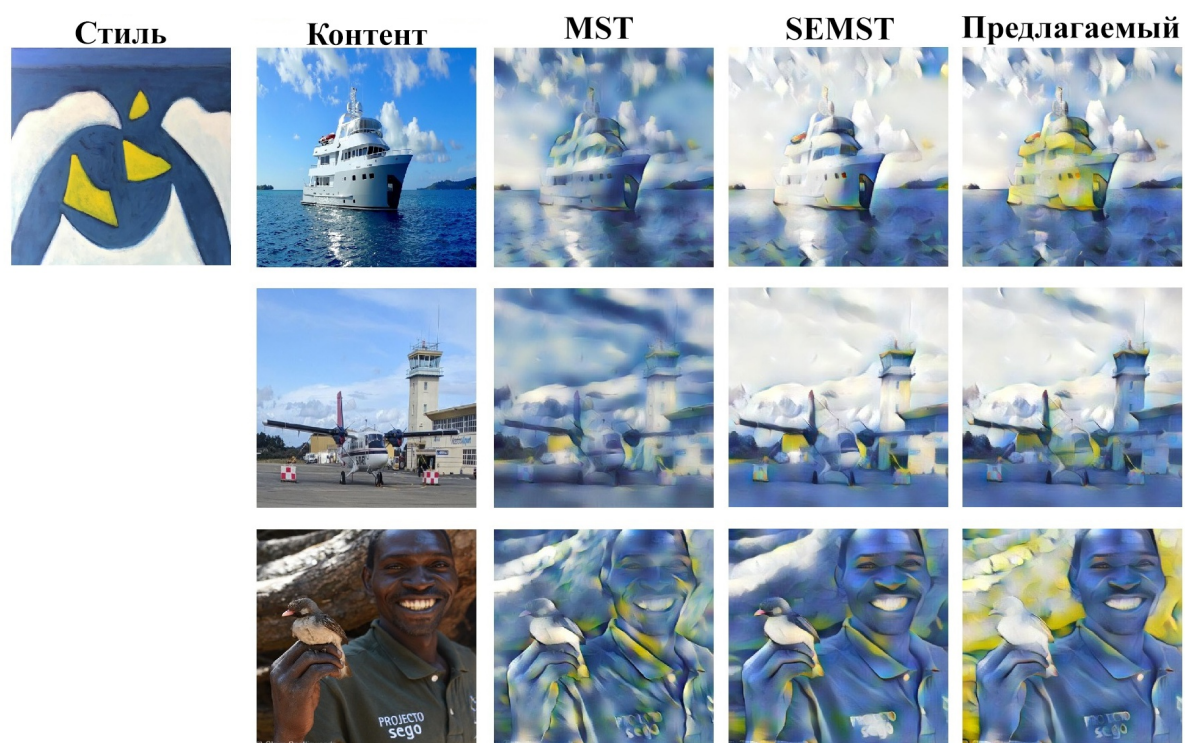


Рис. 7. Сравнение стилизаций с существующими методами мультимодальной стилизации MST [13] и SEMST [14].

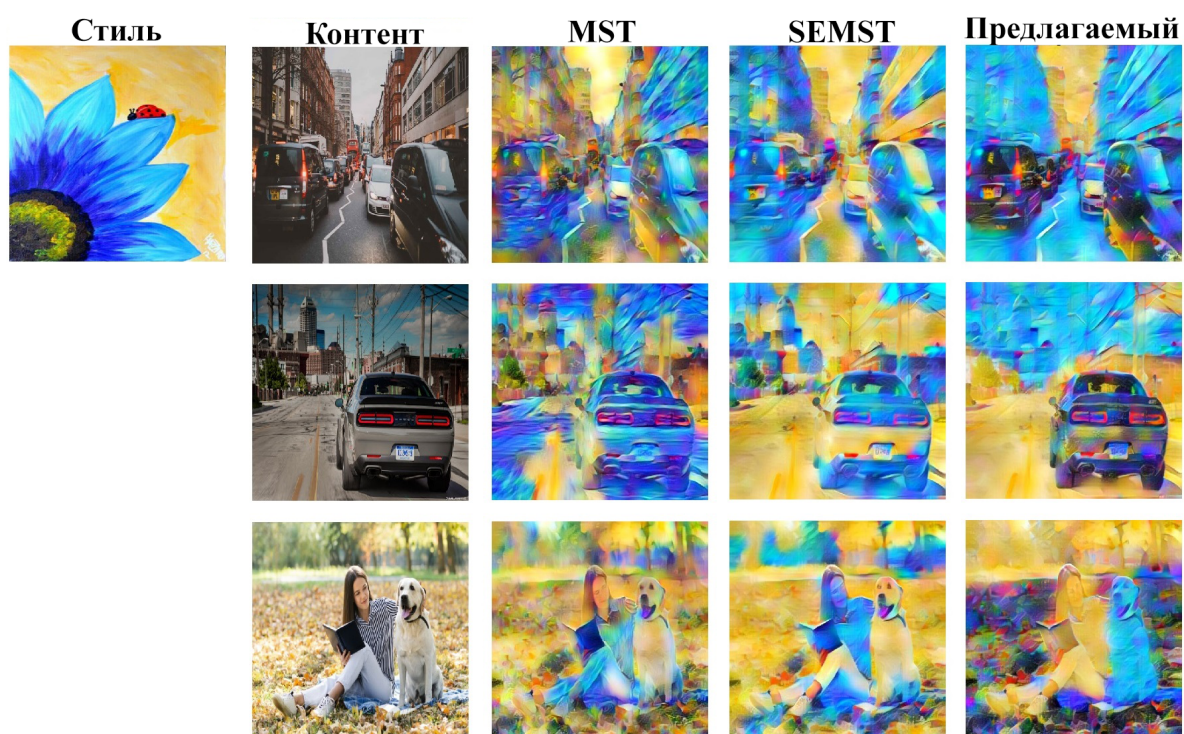


Рис. 8. Сравнение стилизаций с существующими методами мультимодальной стилизации MST [13] и SEMST [14].

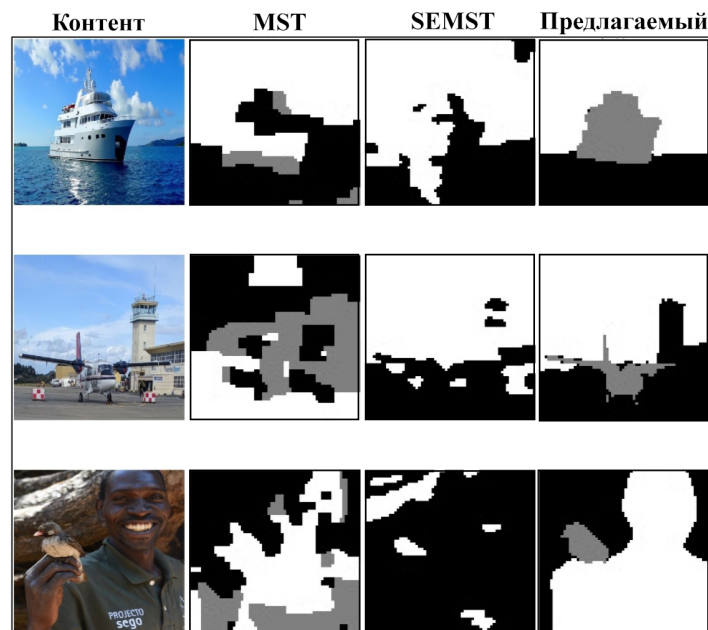


Рис. 9. Сравнение масок сегментации контента с методами-конкурентами.

Результаты сегментации контентного изображения на рис. 9 предлагаемого подхода улучшенной мультимодальной стилизации демонстрируют значительное увеличение в точности по сравнению с алгоритмом кластеризации К-средних, используемым в подходе *SEMST* [14] и, особенно, по сравнению с *MST* [13], в котором контент не был сегментирован/кластеризован, а напрямую участвует в функции поэлементного сопоставления пространственных точек сверточных представлений с центрами стилевых кластеров. Обратим внимание, что алгоритм автоматического подбора числа кластеров, предлагаемый в *SEMST*, как правило, сходится к значению $K = 2$, что доказывает недостаток универсальности и гибкости указанного подхода, поскольку на практике число наблюдаемых стилевых кластеров может быть больше, либо кластеров может не быть совсем.

Результаты кластеризации стилового изображения (рис. 10) предлагаемым в работе алгоритмом демонстрируют видимое улучшение в качестве по сравнению с существующими методами мультимодальной стилизации *MST* [13] и *SEMST* [14]. Метод *SEMST* обладает теми же проблемами, что и кластеризация контента, а метод *MST* теряет точность в выделении областей из-за проведения процедуры кластеризации в глубоком свёрточном слое сети VGG-19, а также не обладает гибкостью своей работы ввиду требований от пользователей в указании числа кластеров K .

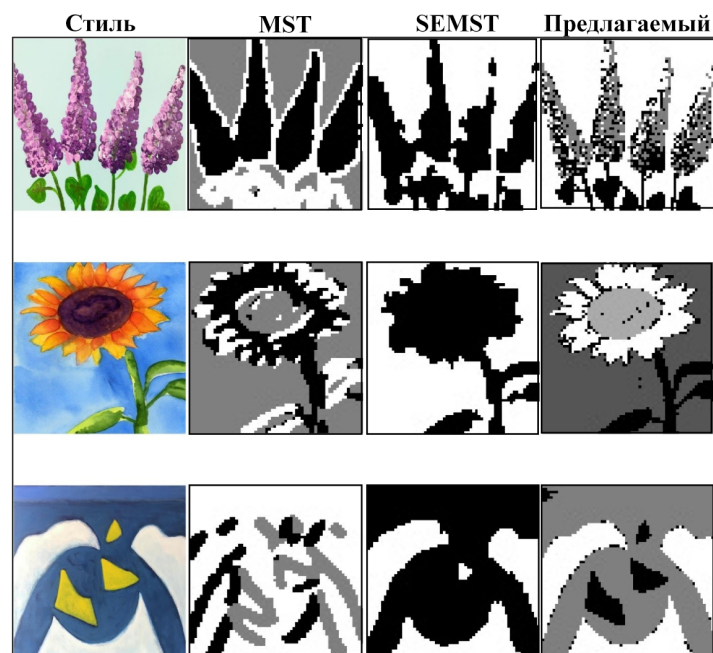


Рис. 10. Сравнение масок кластеризации стиля с методами-конкурентами.

Результаты опроса респондентов

Для того, чтобы сравнить качество стилизаций предложенного метода *IMST* и существующих методов мультимодальной стилизации *MST*, *SEMST* обратимся к опросу независимых респондентов. Для генерации стилизаций было выбрано 22 случайные пары контентного и стилизованных изображений. Единственным критерием отбора стилизованных изображений было требование наличия у них четкой структуры композиции картины, иначе различия между стилизациями получались практически не заметными. Для каждой пары были сгенерированы три варианта стилизации предложенным методом и методами *MST* и *SEMST*. Результаты стилизаций, вместе с исходными контентными и стилизованными изображениями, были показаны 14 респондентам в случайном порядке, и от них требовалось выбрать наиболее удачные из них. Каждый из 14 респондентов должен был сделать 22 выбора в пользу того или иного метода, что приводит к суммарному количеству в 308 отзывов. Респонденты не были знакомы с участвующими в сравнении методами стилизации. Таблица 1 суммаризирует собранные мнения респондентов.

Табл. 1. Голоса респондентов за тот или иной метод.

	MST	SEMST	Предлагаемый
Голоса (количество)	66	87	155
Голоса (проценты)	22%	28%	50%

Исходя из результатов опроса распределение голосов значительно отличается от равномерного и достигает максимального значения у предлагаемого в данной работе метода, что объясняется улучшенным разбиением контентного изображения на сегменты с помощью сегментации, использованием более продвинутого метода кластеризации изображения со стилем и предложенным алгоритмом автоматического сопоставления между сегментами контента и кластерами стиля.

Заключение

Классические методы автоматической стилизации изображений извлекают стиль равномерно со всего стилового изображения и накладывают его также равномерно на все контентное изображение, что снижает выразительность стилизаций и не соответствует рисункам художников, использующим разные особенности стиля для разных объектов на изображении. В работе даны подробные описания как классических алгоритмов стилизации, так и мультимодальных (где стиль для разных участков накладывается по-разному) и представлен новый метод мультимодальной стилизации, в котором предложены новая схема разбиения контентного и стилового изображений, а также оригинальный алгоритм их сопоставления друг с другом.

Предлагаемый алгоритм генерирует более выразительные и насыщенные стилизации, расставляя акценты на основные объекты сцены. По сравнению с существующими методами мультимодальной стилизации предложенный алгоритм генерирует стилизации лучшего качества, что подтверждается сравнительным анализом и результатами опросов респондентов.

Литература

1. *Gooch B., Gooch A.* Non-photorealistic rendering // AK Peters/CRC Press. 2001.
2. *Strothotte T., Schlechtweg S.* Non-photorealistic computer graphics: modeling, rendering, and animation // Morgan Kaufmann. 2002.
3. *Rosin P., Collomosse J.* Image and video-based artistic stylisation // Springer Science & Business Media. **42**. 2012.
4. *Gatys L.A., Ecker A.S., Bethge M.* Image style transfer using convolutional neural networks // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016. P. 2414–2423.
5. *Johnson J., Alahi A., Fei-Fei L.* Perceptual losses for real-time style transfer and super-resolution // European conference on computer vision. Springer, 2016. P. 694–711.
6. *Dumoulin V., Shlens J., Kudlur M.* A learned representation for artistic style // Proc. of ICLR. Springer. **2**. 2017.

7. *Chen D., Yuan L., Liao J., Yu N., Hua G.* Stylebank: An explicit representation for neural image style transfer // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. P. 1897–1906.
8. *Huang X., Belongie S.* Arbitrary style transfer in real-time with adaptive instance normalization // Proceedings of the IEEE International Conference on Computer Vision. 2017. P. 1501–1510.
9. *Ghiasi G., Lee H., Kudlur M., Dumoulin V., Shlens J.* Exploring the structure of a real-time, arbitrary neural artistic stylization network // arXiv preprint arXiv:1705.06830. 2017.
10. *Li Y., Fang C., Yang J., Wang Z., Lu X., Yang M.H.* Universal style transfer via feature transforms // Advances in neural information processing systems. 2017. P. 386–396.
11. *Gatys L.A., Ecker A.S., Bethge M., Hertzmann A., Shechtman E.* Controlling perceptual factors in neural style transfer // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017. P. 3985–3993.
12. *Simonyan K., Zisserman A.* Very deep convolutional networks for large-scale image recognition // arXiv preprint arXiv:1409.1556. 2014.
13. *Zhang Y., Fang C., Wang Z., Lin Z., Fu Y., Yang J.* Multimodal style transfer via graph cuts // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019. P. 5943–5951.
14. *Chen C.* Structure-emphasized Multimodal Style Transfer // Zenodo 10.5281/zenodo.3602064. 2020.
15. *Campello R., Moulavi D., Sander J.* Density-based clustering based on hierarchical density estimates // Pacific-Asia conference on knowledge discovery and data mining. Springer, 2013. P. 160–172.
16. *Chen L.C., Papandreou G., Schroff F., Hartwig A.* Rethinking atrous convolution for semantic image segmentation // arXiv preprint arXiv:1706.05587. 2017.
17. *Ranftl R., Lasinger K., Hafner D., Schindler K., Koltun V.* Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer // IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2020.
18. *Neubert P., Protzel P.* Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms // 2014 22nd international conference on pattern recognition. IEEE, 2014. P. 996–1001.