PeCA: Palette Context Assisted Inference for Test-Time Paint-Bucket Colourisation on Animation Videos

Dongheng Lin and Jianbo Jiao

The MIx Group, University of Birmingham

Abstract. In animation production, paint-bucket colourisation for hand-drawn animation is a labour-intensive procedure that assigns each enclosed region in line sketches a colour from reference design sheets. Recent automatic paint-bucket colourisation pipelines mirror this workflow via region correspondence, but correspondences can be brittle when regions are ambiguous fragments without proper context. In this paper, we propose Palette Context Assisted (PeCA), a new training-free, plug-and-play framework for animation video colourisation that aims to close this gap at test-time via reasoning over spatial and temporal contexts. Extensive experiments on existing benchmarks and a newly introduced long-video test case show consistent performance boosts. [1]

Keywords: Video Colourisation · Animation · Correspondence

"A colour shines in its surroundings."

Ludwig Wittgenstein

1 Introduction

Hand-drawn animation colourisation is not an unconstrained image synthesis problem. In production, region colours must strictly follow a discrete celluloid palette defined by design sheets, and colours must be correctly assigned despite deformation, occlusion, and changes in the layout of line-enclosed regions [25]. Consequently, transforming line sketches into correctly coloured animation remains a major bottleneck [10]. A particularly labour-intensive stage is the paint-bucket colourisation, where artists meticulously assign colours to a massive number of enclosed regions [25]. This step is repetitive yet unforgiving: even minor colourisation mistakes or boundary leaks can break production-level quality and lead to costly correction. Although automatic colourisation has improved markedly with the advancement of generative computer vision techniques [10,41], fully reliable automated paint-bucket colourisation remains a challenge.

[1] Project page: `https://rathgrith.github.io/`

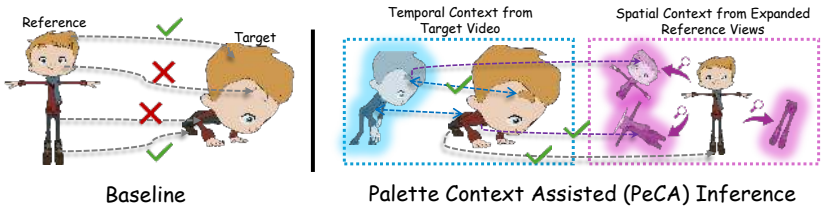


Fig. 1: Conceptual overview of PeCA. Compared to a baseline that colours regions using isolated correspondences, PeCA leverages temporal and spatial context to pick supports for a more reliable colourisation.

Existing automated colourisation approaches can be broadly split into pixel-based generative models and segment-based pipelines. Pixel-generative methods [3, 11, 19, 33, 36, 46, 50, 53], including recent DiT-based generative models [17, 47, 49, 54], can produce seemingly pleasing renders, but often violate production constraints: colours bleed across ink boundaries and discrete colour restriction is weak or absent [41]. Segment-based methods [2,5–7,21,24,30] instead mirror the paint-bucket workflow by treating colourisation as assigning a colour to enclosed regions in the target frame from those in reference frames (*i.e.* coloured key-frames or design sheets) by regional correspondences. This formulation preserves the exact colour and region constraints by construction. The remaining difficulty lies in finding robust correspondences between references and target frame regions that may appear to be completely different.

Most segment-based pipelines solve paint-bucket colourisation by region correspondences, which parse regional similarities to colour label estimations either by top matches or a weighted combination of them [5,7,8,21,24]. Although some methods introduce structural or temporal constraints, the per-region assignment is still largely driven by local correspondence scores. In practice, failures are often not due to a complete absence of correct matches, but arise when correspondences are ambiguous or noisy due to view/pose variations in animation videos [20, 24]. For example, thin fragments under occlusion can be visually ambiguous on their own, and multiple reference regions may look similarly plausible. This ambiguity persists and becomes a bottleneck. As a result, despite being trained on correspondence-based colourisation, such a direct segment matching & colour propagation pipeline still struggles to deal with spurious correspondences under reference-target appearance gaps [20, 24].

A more robust view is that region identity in animation is rarely resolved from a single isolated match. Humans rely on context when understanding colours in art [45]: both from similar views in reference images, and the temporal continuity in target videos themselves. This suggests a complementary direction to training a stronger backbone: when a direct reference-to-target match is uncertain, spatially similar reference views and temporally neighbouring frames can provide indirect context support that refines ambiguous colour propagation.

To this end, we propose the Palette Context Assisted (PeCA) inference framework that improves segment matching colourisation with test-time con-

text, as shown in Fig. 1. PeCA is training-free at inference; it fits well to either trained colourisation models or frozen foundation backbones. It first constructs a target-conditioned support reference bank, so that reference views are expanded to have better coverage of the current target shot, reducing the visual gap with spatially-close context (Sec. 3.3). We then resolve noisy top correspondences with a soft voting scheme (Sec. 3.4) to reach a consensus that blocks spurious correspondences. Finally, we refine per-region assignments across time, leveraging the continuity between adjacent frames while avoiding unreliable transfers via a gated mechanism (Sec. 3.5). Together, PeCA turns noisy, isolated segment colourisation into context-aware colour assignments without task-specific training. We summarise our contributions as follows:

- We propose a plug-and-play Palette Context Assisted (PeCA) inference framework for paint-bucket colourisation, leveraging context at test-time.
- PeCA improves the default inference that struggles with region ambiguity, by aggregating contexts from both spatially-supportive reference views and temporal continuity in animation videos.
- Extensive experiments on existing benchmarks and a newly introduced long-shot test show consistent gains on both task-trained and frozen foundation backbones, with particularly larger improvements in training-free settings.

2 Related Works

2.1 Automated Paint-bucket Colourisation in Animation

Production paint-bucket colourisation assigns each segmented region a colour from a fixed palette, thus fundamentally relies on region correspondence [25]. Early segment-based methods modelled this problem as geometric or graph-based matching between regions [21, 30]. These approaches typically assume moderate motion and rely on handcrafted similarity measures or motion cues, which fail to handle large appearance gaps and longer videos in production [8]. BasicPBC [7] departs from this simple region matching formulation by explicitly modelling topology with inclusion and subset matching to handle split/merge events, and designed propagation strategies tailored to these cases. Feng *et al*. [8] further extended this direction with a unified pipeline that augments adjacent-frame matching with temporal-structural constraints and additional refinement modules. Compared with earlier methods, they incorporate more structured matching rules and model-specific post-processing to refine colour propagation. But still, these works primarily operate under a temporal-local assumption, where the target frame and reference frames are in the same video.

Key-frame colourisation task relaxes this assumption and considers arbitrary reference–target pairs, such as design sheets and distant shots. BasicPBC-Ref [6] adapts the segment-matching framework to this setting by incorporating stronger semantic features to bridge pose and layout gaps. DACoN [24] shows that region matching with powerful foundation model descriptors can already serve as a strong baseline for key-frame colourisation, and further improves performance

through additional training and model components. Despite such progress, these prior works still report failure modes on extreme view/poses that are visually different from reference shots, and the gains by introducing more reference shot become more marginal as the number of references goes up [24]. This saturation revealed a universal bottleneck identified in many tasks, when models fail to *find direct correspondences under huge visual gaps* [12].

As a response to this bottleneck, our PECA builds on the simplest segment-matching formulation. Rather than introducing a stronger model for direct correspondence, we focus on improving robustness from a model-agnostic perspective. By keeping the underlying model unchanged, our approach remains compatible with both trained segment matching models and frozen foundation backbones, and isolates the effect of PECA reasoning from model-specific design choices.

2.2 Visual Correspondence by Foundation Models at Test-time

Large pretrained models provide transferable representations that enable training-free or low-supervision correspondence and region reasoning [14, 26, 28, 31]. A common practice is to pool dense features within regions to build local descriptors, then perform similarity-based retrieval for region matching and propagation [13,34,37]. Beyond such representation reuse, another line of work [16,22,27, 43, 48, 51, 52] improves the performance of various downstream tasks (including generic video colourisation) via test-time adaptation, updating model parameters or refinement at inference steps.

In production-oriented animation paint-bucket colourisation, models must generalise to binarised line sketches with sparse appearance cues. As a result, even strong pretrained region descriptors [26, 28, 31] can be brittle when correspondence is solved purely by similarity retrieval without task-specific training [24]. We thus take a test-time perspective and aim to make inference more reliable with minimal assumptions on the backbone. Instead of treating each region match as an isolated decision, we organise region-level evidence into a palette-space belief and refine it using spatial support from references and temporal support from the target sequence.

3 Methodology

3.1 Problem Formulation

We study production-oriented paint-bucket colourisation for hand-drawn animation, where each enclosed region must be assigned a colour from a discrete palette in reference regions. Given a target video clip of T sketch frames $\{I_t\}_{t=1}^{T}$, each frame is partitioned into closed regions $\mathcal{S}_t = \{s_{t,i}\}_{i=1}^{N_t}$. A reference set $\mathcal{R}$ is provided as $\mathcal{R} = \{(I^{(r)}, \mathcal{S}^{(r)}, Y^{(r)})\}_{r=1}^{R}$, where $Y^{(r)}$ assigns a ground truth colour label (from a finite palette $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$) to each reference region in $\mathcal{S}^{(r)}$ (segmented from line-sketches $I^{(r)}$ by flood fill [40]). Our goal is to assign each target region $s_{t,i}$ a colour label $\hat{y}_{t,i} \in \mathcal{C}$ correctly throughout the video.

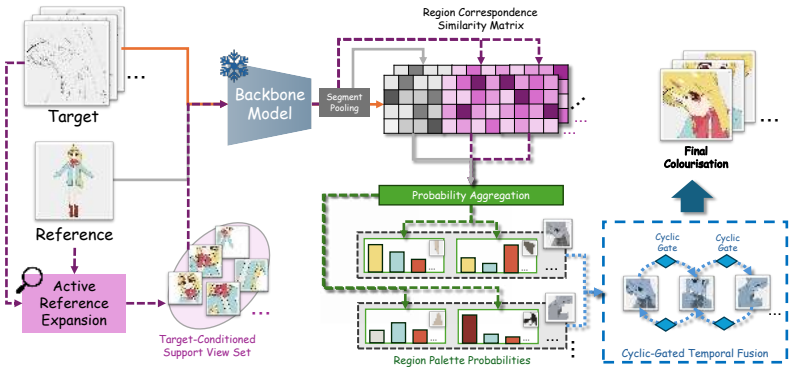


Fig. 2: The proposed PeCA framework overview. Active Reference Expansion (Sec. 3.3) builds a target-conditioned reference support set. **Probability Aggregation** (Sec. 3.4) aggregates noisy matches into per-region palette colour probabilities via soft voting. **Cyclic-gated Temporal Fusion** (Sec. 3.5) fuses colour probabilities across adjacent frames through cycle-consistent temporal links, altogether improving colourisation with **spatial**, **probabilistic**, and **temporal** context.

Following segment-matching pipelines [24, 34], we compute a descriptor for each region by average pooling dense features inside its mask in Eq. (1), and define cosine similarity $S_{(t,i),(r,j)} = \langle \bar{\mathbf{f}}_{t,i}, \bar{\mathbf{f}}_j^{(r)} \rangle$. A retrieval-style baseline in previous SOTA [24] copies the colour label from the top match as in Eq. (2):

$$\mathbf{f}_{t,i} = \mathrm{AvgPool}\big(\{\phi(I_t)[p] \mid p \in s_{t,i}\}\big), \qquad \mathbf{f}_j^{(r)} = \mathrm{AvgPool}\big(\{\phi(I^{(r)})[p] \mid p \in s_j^{(r)}\}\big), \quad (1)$$

$$(r^*, j^*) = \arg\max_{r,j} S_{(t,i),(r,j)}, \qquad \hat{y}_{t,i} = y_{j^*}^{(r^*)}. \quad (2)$$

As discussed in Sec. 1 and Sec. 2, direct correspondence probability $S_{(t,i),(r,j)}$ can be ambiguous for colour propagation. This motivates us to propose the Palette Context Assisted (PeCA) test-time reasoning framework that exploits spatial and temporal context of correspondences and colours during inference.

3.2 Palette Context Assisted (PeCA) Framework Overview

We propose PeCA, a training-free and plug-and-play inference framework that improves region matching-based colourisation by constructing and exploiting context at test time (Fig. 2). PeCA first builds a target-conditioned support bank by expanding the given reference shots to a limited support set that maximises spatial coverage to the target video (Sec. 3.3). Then, from the multi-source correspondences, a soft top-k voting converts multiple plausible matches into per-region palette-colour consensus probabilities (Sec. 3.4). Finally, we refine these probabilities along reliable correspondence links between adjacent frames serving as temporal context (Sec. 3.5).

3.3 Spatially-Supportive Reference Views Selection

Prior works [6, 24] have shown that increasing the number and diversity of reference shots brings consistent gains for segment-matching-based colourisation, largely because it improves the chance that a target region finds a reference region from a similar layout. In real productions, however, we are often given limited reference RGB images due to the labour-intensive nature of colourisation. In this regard, a natural next step is to expand this limited reference bank with test-time augmentation [35], generating new views to reference shots that possibly provide easier "shortcut" matchings for colour propagation.

However, naive test-time augmentation is unlikely to scale efficiently at inference time under this setting [38]. In basic inference pipelines [5, 24], region assignment is resolved from a similarity matrix over all reference-region candidates. Naively augmenting views, which increases the candidate set indiscriminately, may provide useful evidence. But the additional views may also increase exposure to high-similarity distractors [29]. This makes top matches less stable for ambiguous regions, resulting in a limited performance gain (see Tab. 5). This also aligns with a trend in previous works where further adding more references (>5) shows marginal gains [24].

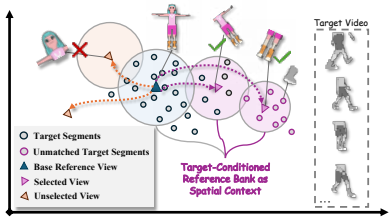


Fig. 3: Active Reference Expansion. Target region features are shown in gray (hard regions in purple). We aim to select augmented views that maximise coverage to targets while excluding views that provide limited support to target video.

We therefore designed active reference expansion (Fig. 3) to make additional reference views both budgeted and target-aware, forming a spatial context that best covers the target regions. Starting from the original reference views $\mathcal{V}_0$, we first generate an augmented candidate pool $\mathcal{V}_{\text{aug}}$ by applying geometric transformations (flips, rotations, affine transforms; details in the supplementary material). Rather than keeping all candidates, we select only B views that best support the target video in the feature space, so the bank is strengthened without indiscriminately enlarging the region candidate set. Specifically, for each candidate view $v \in \mathcal{V}_{\text{aug}}$ with region descriptors $\{\mathbf{f}_j^{(v)}\}_{j=1}^{N_v}$, we measure its support to a target region (t, i) by the similarity of its best-matching region:

$$\text{score}_v(t,i) \;=\; \max_{1\le j\le N_v} \left\langle \bar{\mathbf{f}}_{t,i},\, \bar{\mathbf{f}}_j^{(v)} \right\rangle. \tag{3}$$

We expect the expansion to keep a subset $V \subseteq \mathcal{V}_{\text{aug}}$, such that the resulting support for (t, i) is $\max_{v\in V} \text{score}_v(t, i)$, *i.e.* the best support offered by selected views. We therefore choose B views by maximising a facility-location objective over target regions in uniformly sampled frames $\mathcal{T}_s \subset \{1, \ldots, T\}$:

$$F(V) \;=\; \sum_{t\in\mathcal{T}_s} \sum_{i=1}^{N_t} \max_{v\in V} \text{score}_v(t,i), \qquad \max_{V\subseteq\mathcal{V}_{\text{aug}}} F(V) \text{ s.t. } |V| = B. \tag{4}$$

In practice, we form an augmented candidate pool of size $|\mathcal{V}_{\text{aug}}| = mB$ and select B views from it. This selection is performed once per target video, with overhead depending only on mB (see supplementary for details). Since $F(V)$ is monotone submodular, a greedy algorithm provides a standard approximation guarantee [15]. The resulting support bank $\mathcal{V}_0 \cup V$ improves target-shot coverage under a fixed budget by actively prioritising views that best support the current video. Importantly, this selection step controls the exposure to distractor regions, yielding a more relevant candidate set for subsequent matching. On the other hand, with more reference shots, naturally, for each target region, it introduces additional correspondence candidates from different views. This motivates our next step, which votes multiple high-confidence correspondences to palette colour probabilities (Sec. 3.4).

3.4 Correspondence Candidate Voting for Colour Palette

When more than one reference shots are available, a target region is naturally supported by multiple high-confidence candidates from different sources. To better utilise such probabilistic context, instead of committing to top-1 retrieved region [24] or full linear combinations [5], we wish to resolve the correspondences to colour estimations by *soft-voting* over the top-k correspondence hypotheses. This produces per-region colour consensus probability over the finite palette from the relevant correspondence probabilistic context.

Specifically, for each target region (t, i), let $\mathcal{N}_k(t, i)$ denote the top-k candidate reference regions under similarity $S_{(t,i),(r,j)}$. We convert similarities into normalised weights with temperature $\tau \in (0, 1]$:

$$p_{t,i}(r, j) = \frac{\exp\left(S_{(t,i),(r,j)}/\tau\right)}{\sum_{(r',j')\in\mathcal{N}_k(t,i)} \exp\left(S_{(t,i),(r',j')}/\tau\right)}, \qquad (r, j) \in \mathcal{N}_k(t, i). \tag{5}$$

We then vote these weighted matches into palette colour space:

$$P_{t,i}(c) = \sum_{(r,j)\in\mathcal{N}_k(t,i)} p_{t,i}(r, j) \cdot \mathbb{1}\left[y_j^{(r)} = c\right], \qquad c \in \mathcal{C}. \tag{6}$$

Here, $P_{t,i} \in \Delta^{|\mathcal{C}|}$ summarises the correspondence evidence as colour-level probabilities, where $\Delta^{|\mathcal{C}|}$ denotes the probability simplex over $|\mathcal{C}|$ colour entries. The temperature τ controls the sharpness of this soft vote. When k=1 and $\tau \to 0$, it reduces to hard copying as in baseline, while moderately larger k or τ pool evidence across matches and reduce sensitivity to spurious correspondence.

Serving as an interface from region correspondence to per-region colourisation results, the aggregation has two practical benefits for spatial/temporal context: First, restricting aggregation to top-k candidates avoids the dilution effect of naive global mixing when the candidate pool gets larger after active expansion in Sec. 3.3, leading to robust colourisations from soft voting by multiple candidates. Second, colour probabilities live in a fixed label simplex shared by all frames,

making them directly comparable and therefore suitable for supporting each other as temporal contexts in Sec. 3.5, unlike matching probabilities that may change across regions with the same colour identity but in different frame pairs.

3.5 Temporal Cycle-Consistency for Colour Refinement

Now that each frame has per-region palette-colour probabilities constructed from a strengthened reference support. In addition, animation videos also provide a temporal cue, in which adjacent frames within the same video look more similar due to temporal continuity of videos [4, 41, 44]. Motivated by this, if a direct match between reference and target failed, an easier transitive colour-propagation shortcut can be parsed from the adjacent frames' correspondences (a temporal context) to refine the colourisation.

However, region matching is not always reliable under changes between adjacent frames. If we indiscriminately fuse information across time, a single spurious match can be propagated to other frames and amplified. We therefore refine colour probabilities only along cycle-consistent temporal links, so that temporal context is used when it is reliable and ignored otherwise, as shown in Fig. 4.

Specifically, between adjacent frames, we compute adjacent-frame region similarity $A_t[i,j] = \langle \bar{\mathbf{f}}_{t,i}, \bar{\mathbf{f}}_{t-1,j} \rangle$. We define the forward nearest-neighbour match from frame t to $t-1$ as $\pi_t(i) = \arg\max_j A_t[i,j], \quad i \in \{1,\dots,N_t\}$, and symmetrically, the backward match from frame $t-1$ to t as $\rho_t(j) = \arg\max_i A_t[i,j], \quad j \in \{1,\dots,N_{t-1}\}$. We treat a temporal link as cyclic stable only if it is bidirectional:

$$(t,i) \text{ is stable if } \rho_t\big(\pi_t(i)\big) = i. \tag{7}$$

This cycle-check conservatively filters unreliable correspondences, which is crucial as the matching between adjacent frames can still be noisy (see Tab. 6). For cyclic stable links, we fuse colour probabilities with a product update:

$$\tilde{P}_{t,i}(c) \propto P_{t,i}(c) \cdot P_{t-1,\pi_t(i)}(c), \qquad P_{t,i} \leftarrow \text{Normalise}(\tilde{P}_{t,i}), \quad c \in \mathcal{C}. \tag{8}$$

Intuitively, Eq. (8) reinforces colour propagation for hard cases by leveraging palette probabilities carried by regions in other frames of the same video, while a cycle-consistency gate prevents error propagation. We perform one forward sweep ($t = 2 \to T$) and one backward sweep ($t = T-1 \to 1$). In the backward sweep, we apply the same matching, cycle check, and fusion with indices swapped. In practice, the two passes are complementary: each direction conditions on different neighbouring frames that may already have more reliable estimates, so bidirectional sweeping improves robustness and yields more stable colourisation.

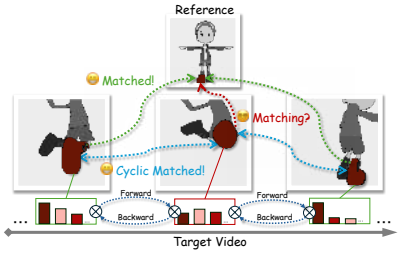


Fig. 4: Cyclic-Gated Temporal Fusion utilises temporal context by fusing per-region colour probabilities only along cycle-consistent matches between adjacent frames, refining colours while avoiding unreliable temporal fusions.

4 Experiments

4.1 Experiment Setup

We evaluate on the existing PaintBucket benchmarks [6, 7] sourced from CG-Rendered and hand-drawn videos. We further test our method against previous works on a newly constructed long-shot dataset, with 10–20× the length of target video, compared to previous datasets. Further experiments of segment matching on generic video dataset and ablations are provided in supplementary material.

PaintBucket-Character (PBC-3D) dataset contains 22 characters with 9–16 reference design-sheets per character; following prior work [6,24], we report results on the official test split of 3,000 frames.

PaintBucket-Real (PBC-Real) is a hand-drawn test set collected from professional animation, with 200 frames in total (20 short clips). Since it is not provided with design-sheet references, we use the first coloured frame from the clips as references, following previous works [7, 8, 24].

Anita-Pirate. We annotated a new long-shot benchmark featuring a hand-drawn 206-frame sequence, with raw data from the Anita dataset [1], which is more challenging than PBC-Real due to its 10–20× longer time horizon with roughly 140 segments per frame. *I.e.*, it comes with frames with complex region layouts over a long time horizon. We will release this new challenging test set. Annotation protocol and licensing details are provided in the supplementary.

Evaluation protocols: We consider several production-relevant settings. Firstly, we evaluate standard design-sheet key-frame colourisation, following previous works [6,24], in Tab. 1. This setting provides a few coloured design sheets without assuming any temporal relation to the target video, and is therefore the most general reference protocol [6, 24]. We also evaluate video colourisation under the same-video key-frame colourisation, in which references are sampled from the target video itself. Note that this is the only reference type available for PBC-Real and Anita-Pirate without design sheets. We report two protocols from prior works. 1) First-frame reference (Tab. 3): only the first frame is given as an RGB reference [24]. 2) Two-sided reference (in-between): only the first and last frames are given as RGB references [8] (Tab. 4). Following previous works [6–8, 24], we report both segment- and pixel-level metrics: `Acc`, `Acc-Thresh` (segments>10 pixels), `Pix-Acc`, `Pix-F-Acc` (for foreground), and `Pix-B-MIoU` (for background). All metrics are reported in percentages, and larger values mean better performance. See further details in the supplementary.

Implementation details: We compare against representative pixel-generative methods [19, 23, 53] and segment-based pipelines [6, 7, 24] under official protocols [6, 24]. We build a *Base* inference for SOTA [24] and frozen foundation models that performs colourisation using region correspondence just as previous works did. We then apply PeCA as a plug-and-play inference framework to various backbone models, either with or without colourisation task training, and fixed hyperparameters (top-k=64, τ=0.05 and #views $B = 31$, m=4.) across all experiments. Further implementation details, computational costs, and hyperparameter settings are provided in the supplementary material.

Table 1: One-shot key-frame (design-sheet) colourisation on PBC-3D. We follow prior work and use a single design-sheet reference image to colourise the target video. We compared the **PECA** framework on both the trained DACoN 1.1 Model and other frozen foundation models with the **base** setting that matches the segmented region features for prediction. **Training-free** indicates whether the backbone is used without further training (✓) or requires training (✗) on colourisation tasks.

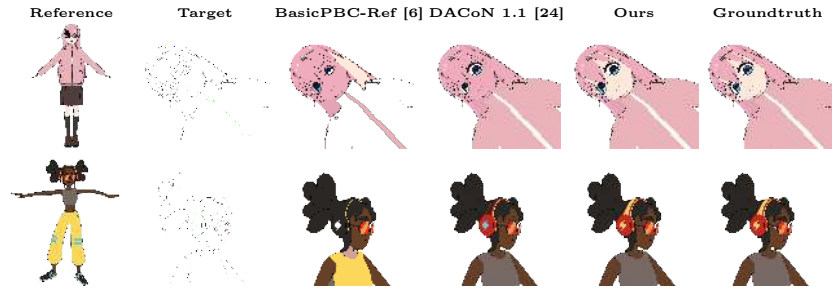
Method / Backbone	Training-free	Acc (%)	Acc-Thresh (%)	Pix-Acc (%)	Pix-F-Acc (%)	Pix-B-MIoU (%)
ColorFlow [53]	✗	9.72	10.81	50.64	9.16	57.17
MangaNinja [19]	✗	14.86	16.73	7.11	28.52	0.00
AniDoc [23]	✗	19.80	22.68	77.38	46.46	87.32
Cobra [54]	✗	15.06	17.26	69.20	19.72	82.69
MagicColor [49]	✗	21.48	24.81	16.34	44.04	7.63
BasicPBC-Ref [6]	✗	52.55	56.73	90.53	72.33	94.56
DACoN [24]	✗	67.87	72.58	96.99	91.00	99.08
DACoN 1.1 [24]	✗	68.01	72.87	96.97	91.03	99.11
DACoN 1.1 + PECA	✗	**72.04** (4.03↑)	**77.08** (4.21↑)	**97.90** (0.93↑)	**94.04** (3.01↑)	**99.42** (0.31↑)
SAM2.1-Large (Base) [31]	✓	34.54	38.95	86.76	54.12	88.37
SAM2.1-Large + PECA	✓	**46.65** (12.11↑)	**49.92** (10.97↑)	**88.70** (1.94↑)	**66.96** (12.84↑)	**96.70** (8.33↑)
DINOv3 ConvNeXT-L (Base) [39]	✓	34.90	36.35	71.32	49.79	75.93
DINOv3 ConvNeXT-L + PECA	✓	**45.88** (10.98↑)	**46.97** (10.62↑)	**80.13** (8.81↑)	**60.15** (10.36↑)	**85.38** (9.45↑)
SigLIPv2 ViT-B/16 (Base) [42]	✓	48.64	51.68	89.24	70.05	91.03
SigLIPv2 ViT-B/16 + PECA	✓	**55.34** (6.70↑)	**58.88** (7.20↑)	**92.48** (3.24↑)	**80.37** (10.32↑)	**93.88** (2.85↑)
DINOv2 ViT-L/14 (Base) [26]	✓	57.49	61.86	95.35	87.24	97.45
DINOv2 ViT-L/14 + PECA	✓	**61.38** (3.89↑)	**65.58** (3.72↑)	**96.25** (0.90↑)	**89.31** (2.07↑)	**98.62** (1.17↑)

Table 2: Key-frame (design-sheet) colourisation on PBC-3D with more references. We report results under the multi-reference protocols from [6,24] for methods requires colourisation training (✗) or not (✓).

# of Refs	Method / Backbone	Training-free	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
5-shot refs	ColorFlow [53]	✗	12.64	14.37	54.51	15.26	61.22
	BasicPBC-Ref [6]	✗	–	64.59	96.12	83.17	98.67
	DACoN [24]	✗	73.25	77.44	97.74	93.70	99.13
	DACoN 1.1 [24]	✗	73.91	78.23	97.84	94.28	98.92
	DACoN 1.1 + PECA	✗	**77.73** (3.82↑)	**82.39** (4.16↑)	**98.87** (1.03↑)	**97.02** (2.74↑)	**99.45** (0.53↑)
	SAM2.1-Large (Base) [31]	✓	43.80	46.59	87.66	62.25	96.75
	SAM2.1-Large + PECA	✓	**57.23** (13.43↑)	**60.96** (14.37↑)	**91.50** (3.84↑)	**76.52** (14.27↑)	**97.18** (0.43↑)
	DINOv2 ViT-L/14 (Base) [26]	✓	62.65	66.42	96.77	91.54	97.96
	DINOv2 ViT-L/14 + PECA	✓	**66.46** (3.81↑)	**70.01** (3.59↑)	**97.73** (0.96↑)	**93.57** (2.03↑)	**98.83** (0.87↑)
max-shot refs	DACoN [24]	✗	74.31	78.48	98.04	94.27	99.10
	DACoN 1.1 [24]	✗	75.05	79.23	98.19	94.79	99.16
	DACoN 1.1 + PECA	✗	**79.03** (3.98↑)	**83.43** (4.20↑)	**99.01** (0.82↑)	**97.21** (2.42↑)	**99.55** (0.39↑)
	SAM2.1-Large (Base) [31]	✓	46.40	49.30	87.98	63.27	96.59
	SAM2.1-Large + PECA	✓	**56.88** (10.48↑)	**60.50** (11.20↑)	**91.94** (3.96↑)	**77.49** (14.22↑)	**97.29** (0.70↑)
	DINOv2 ViT-L/14 (Base) [26]	✓	63.84	67.67	97.07	91.70	98.28
	DINOv2 ViT-L/14 + PECA	✓	**67.28** (3.44↑)	**70.82** (3.15↑)	**97.71** (0.64↑)	**93.63** (1.93↑)	**98.59** (0.31↑)

4.2 Main Experimental Results

Results on design-sheet key-frame colourisation. Tab. 1 reports one-shot key-frame results with design-sheet references on PBC-3D. On task-trained models, the Palette Context Assisted (PECA) framework further improves the current SOTA on all metrics consistently. Notably, PECA yields substantially larger gains on training-free backbones, indicating that PECA's test-time context reasoning unlocks region matching animation colourisation even for foundation models without colourisation training. Such gain also persists when more key-frame references are given, as shown in Tab. 2, which reports 5-shot and max-shot results on PBC-3D. This further shows that more reference shots do not dilute PECA's contribution to the task. Fig. 5a shows representative qualitative results, showing our method can help the model overcome a huge appearance gap between references and target frames when previous base inference failed to.



(a) Key-frame (Design-sheet) Reference Colourisation Results

(b) First-frame Reference Colourisation Results

Fig. 5: Qualitative comparison under two production formulations. We show one-shot reference and target line sketch, followed by results from **BasicPBC(-Ref)** [6, 7],**DACoN 1.1** [24], our **PECA** on DACoN 1.1, and colour ground-truth (right).

Results on same-video key-frame colourisation. Tab. 3 reports colourisation results on PBC-3D and PBC-Real using only the first frame as a reference to colourise the rest of the video. Including PECA inference provides consistent improvements on the trained DACoN 1.1 pipeline on both domains, and again yields larger gains on training-free backbones. For in-between colourisation, Tab. 4 shows that our PECA consistently improves multiple backbones on both PBC-3D (20-frame clips) and the new long-shot Anita-Pirate dataset. In the latter, only the first and last reference frames are provided with colours, while all the remaining 204 frames have to be coloured, given such a limited reference. All these suggest that test-time context reasoning with PECA can substantially strengthen existing models' performance with different model types and supervisions. Such improvements persist when we naturally have better feature coverage by the temporal proximity of reference-target frames. Corresponding qualitative comparisons are provided in Fig. 5b and Fig. 6.

4.3 Ablations and Analysis

We analyse the three steps in the Palette Context Assisted (PECA) inference framework on both task-trained and frozen foundation backbone under different key-frame reference types. Quantitative results are shown in Tab. 5.

Table 3: First-frame Colourisation on PBC-3D and PBC-Real. We include both colourisation-trained methods (✗) and results using frozen backbones (✓).

Method / Backbone	Training-free	PBC-3D					PBC-Real				
		Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
BasicPBC [7]	✗	56.28	60.14	93.00	77.25	97.19	59.31	62.00	91.84	72.50	98.39
BasicPBC [7] (Online*)	✗	53.18	58.28	93.57	79.92	96.19	57.28	60.47	92.74	74.92	98.35
DAConN [24]	✗	69.91	73.59	97.30	-	-	65.85	69.15	93.50	-	-
DAConN 1.1 [24]	✗	70.34	74.04	97.30	91.13	99.17	65.82	69.11	94.18	80.68	98.76
Nano Banana 2 [9]	✗	–	–	–	–	–	47.78	52.17	90.39	71.63	98.46
DAConN 1.1 + PeCA	✗	**74.41**	**78.08**	**98.11**	**94.06**	**99.50**	**67.64**	**71.29**	**94.70**	**82.11**	**99.48**
StableDiffusion 2.1 (Base) [32]	✓	32.93	34.52	87.38	58.70	94.40	46.45	48.84	89.91	64.13	97.96
StableDiffusion 2.1 + PeCA	✓	**40.50**	**42.01**	**90.87**	**71.01**	**96.51**	**48.11**	**49.70**	**90.89**	**67.45**	**98.18**
SAM2.1-Large (Base) [31]	✓	49.10	52.46	91.64	72.40	97.38	55.63	58.31	90.32	69.21	98.73
SAM2.1-Large + PeCA	✓	**58.98**	**62.89**	**93.65**	**79.72**	**98.11**	**60.41**	**63.44**	**93.25**	**75.99**	**99.00**

* Online setting: the first frame uses the ground-truth reference, and each subsequent frame is colourised using the previous frame's prediction as the reference.

Fig. 6: Qualitative results of in-between colourisation on Anita-Pirate. It compares how our proposed PeCA performs against existing methods when dealing with a 10× longer target frame sequence.

Three steps are jointly contributing. Across both reference types and both backbones, we observe a consistent pattern. Active Reference Expansion (ARE) only gives a minor performance improvement. That means expanding the reference bank, even if in a target-aware way, can still lead to confusion, as it may dilute the exact match with more region candidates. Therefore, further adding Probability Aggregation (PA) yields substantial gains by soft-voting across multiple plausible matches, which reduces sensitivity to spurious matches. Cyclic-gated Temporal-fusion (CT) then provides additional improvements by exploiting temporal context inside the video, refining per-region colour probabilities. Combining all steps gives the strongest results, while removing any of them results in a considerable performance drop, indicating these context cues are complementary and reciprocal as expected, instead of isolated heuristics.

Step-wise contributions differ by reference types. The relative contributions of ARE and CT differ between the two reference types. Under design-sheet references, errors are often driven by spatial mismatch because design sheets can be far from the target shot in pose and region layout; correspondingly, ARE tends to contribute more by improving target-conditioned reference support. Under first-frame references, the reference comes from the same video, and the appearance gap is smaller, so temporal continuity becomes a stronger cue; CT therefore tends to provide more noticeable gains. The balance also depends on the backbone: frozen features benefit more from improved support and aggrega-

Table 4: In-between colourisation results. We report results on short video clips on both the existing PBC-3D [7] testset (20 frames) and a new challenging testset, Anita-Pirate, with 10× longer frame sequence containing multiple complex subjects.

Method/Backbone	Training-free	PBC-3D					Anita-Pirate				
		Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
BasicPBC [7]	✗	63.38	67.77	94.84	84.20	97.54	28.54	28.97	88.52	39.77	96.63
BasicPBC [7] (Online*)	✗	53.97	59.13	93.74	80.62	96.32	7.71	7.94	32.97	17.00	35.93
DACoN 1.1 [24]	✗	78.02	82.11	98.48	95.51	99.47	38.16	39.36	94.29	61.65	99.16
DACoN 1.1 + PeCA	✗	**80.80**	**84.82**	**99.00**	**97.18**	**99.58**	**41.24**	**42.16**	**94.29**	**62.78**	**99.43**
DINOv2 ViT-L/14 (Base) [26]	✓	66.25	70.17	97.73	93.36	98.89	28.55	29.30	93.06	53.88	99.40
DINOv2 ViT-L/14 [26] + PeCA	✓	**69.29**	**72.60**	**98.23**	**94.49**	**99.33**	**31.01**	**31.81**	**93.39**	**57.18**	**99.49**

* Online setting: the first frame uses the ground-truth reference, and each subsequent frame is colourised using the previous frame's prediction as the reference.

Table 5: Ablation study on colourisation under different reference and model types. DACoN 1.1 is pretrained on colourisation; DINOv2 (ViT-L/14) is used frozen in a zero-shot manner. ARE: Active Reference Expansion in Sec. 3.3, PA: Probability Aggregation in Sec. 3.4, CT: Cyclic-gated Temporal-fusion in Sec. 3.5.

Backbone	ARE	PA	CT	Design-sheet (one-shot)					First-frame (one-shot)				
				Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
DACoN 1.1	✗	✗	✗	68.01	72.87	96.97	91.03	99.11	70.34	74.04	97.30	91.13	99.17
	✓	✗	✗	69.04	74.02	97.22	91.67	99.24	70.65	74.54	97.29	91.34	99.25
	✓	✓	✗	70.61	75.40	97.43	92.52	99.28	72.50	76.17	97.64	92.09	99.38
	✗	✓	✓	70.30	75.27	97.21	92.47	99.38	72.78	76.56	97.83	92.87	99.35
	✓	✗	✓	69.02	74.07	97.19	91.68	99.26	70.87	74.80	97.37	91.44	99.31
	✓	✓	✓	**72.04**	**77.08**	**97.90**	**94.04**	**99.42**	**74.41**	**78.08**	**98.11**	**94.06**	**99.50**
DINOv2 ViT-L/14	✗	✗	✗	57.49	61.86	95.35	87.24	97.45	59.50	62.99	96.25	87.95	98.47
	✓	✗	✗	58.74	63.43	95.96	88.14	98.57	60.92	64.53	96.52	88.87	98.64
	✓	✓	✗	60.64	64.66	95.94	88.32	98.60	62.51	65.61	96.80	89.00	99.06
	✗	✓	✓	58.68	62.35	94.56	84.52	97.99	61.00	63.77	96.54	88.76	99.07
	✓	✗	✓	58.51	63.29	95.89	88.08	98.52	61.16	64.85	96.70	89.10	98.97
	✓	✓	✓	**61.38**	**65.58**	**96.25**	**89.31**	**98.62**	**63.28**	**66.47**	**97.13**	**90.45**	**99.22**

tion, whereas the task-trained model benefits more from temporal refinement, consistent with its stronger correspondence quality.

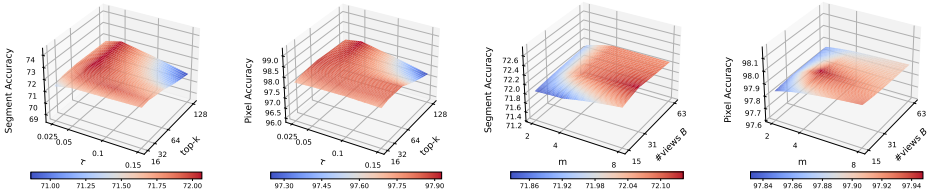
Ablation on view selection and cyclic gating. We additionally ablate two module-internal designs in Tab. 6. For ARE, greedy facility-location selection (Eq. (4)) consistently outperforms keeping all the random views under the same budget, supporting that the gain comes from coverage-aware selection rather than merely adding more augmented views. For CT, disabling the cyclic gate (Eq. (7)) leads to a substantial performance drop, indicating that gating is necessary to prevent unreliable temporal matches from propagating errors over time.

Hyperparameter Sensitivity. Fig. 7a and Fig. 7b visualise the sensitivity of PA (top-k, τ) and ARE (#views B, exploration factor m) under key-frame colourisation. Across a wide range of top-k and τ, all metrics vary mildly, indicating stable behaviour. The observed degradations match expected trends. As $\tau \to 0$ and top-$k \to 1$, probability aggregation degenerates to simple nearest-neighbour [24]. Increasing top-k and τ makes aggregation less selective and approaches a global mixing of candidates, similar in spirit to linear combination schemes used in prior work [5], which can overly flatten the colour distribution and let low-quality matches dilute the prediction. Thus, we select the default setting (τ=0.05, top-k=64) that retains multiple plausible matches while keeping the palette colour probability mass concentrated on high-confidence candidates.

Meanwhile, Fig. 7c and Fig. 7d show that in ARE, the number of selected views B trades off coverage against cost, while m controls the size of the explored

Table 6: Additional one-shot key-frame (design-sheet) colourisation ablations on PECA's internal designs. ARE Selection uses ✗=*random*, ✓=*greedy* (Eq. (4)); CT Cycle Gate uses ✗=*off*, ✓=*on* (Eq. (7)).

Backbone	Selection	Cycle Gate	Acc	Acc-Thresh	Pix-Acc	Pix-F-Acc	Pix-B-MIoU
SAM2.1-Large [31]	✗	✓	47.90	51.41	87.60	63.84	96.74
	✓	✗	45.92	49.11	88.19	65.49	97.18
	✓	✓	**50.69**	**54.59**	**89.10**	**69.05**	**97.39**
CLIP ViT-L/14 [28]	✗	✓	46.45	48.70	88.61	68.66	90.44
	✓	✗	40.30	41.83	88.64	67.45	91.37
	✓	✓	**48.58**	**50.94**	**90.51**	**73.77**	**92.19**



(a) Segment-Acc (top-k vs. τ). **(b)** Pixel-Acc (top-k vs. τ). **(c)** Segment-Acc (#views B vs. m). **(d)** Pixel-Acc (#views B vs. m).

Fig. 7: Hyperparameter sensitivity for PECA inference under key-frame (one-shot design-sheet referenced) colourisation. We visualise the sensitivity of (top-k, τ) in PA (Sec. 3.4) and (#views B vs. m) in ARE Sec. 3.3, reported under Segment Accuracy and Pixel Accuracy.

candidate pool (mB) for greedy selection. Performance saturates quickly as B increases, and increasing m yields only marginal gains beyond moderate search breadth. We therefore adopt B=31 and m=4 as a cost-effective operating point.

5 Conclusion

To sum up, we introduced a training-free and plug-and-play Palette Context Assisted (PECA) inference framework for production-level paint-bucket colourisation, where each region must follow a correct colour in a predefined palette. PECA improves region matching by constructing and validating test-time context: it strengthens spatial context with a target-aware reference expansion, unifying noisy matching evidence in colour space, and uses adjacent frames as temporal context to support prediction. In particular, such context can provide additional support for colour propagation paths when direct matches are ambiguous. Experiments on existing benchmarks and a new long-video test case showed consistent gains on both task-trained models and frozen backbones.

Limitations. Similar to prior works that depend on region segmentation quality and a well-defined palette, PECA cannot guarantee performance under line leakage and missing colours/surfaces in reference (more discussions and failure case analysis please see supplementary material). Future work includes learning stronger reference generation via other generative model [18] or external retrieval beyond simple geometric transformations, and extending test-time context reasoning to interactive production toolkits or pipelines.

References

1. Anita dataset. https://zhenglinpan.github.io/AnitaDataset_homepage/, accessed: 2024-06-24
2. Cao, R., Mo, H., Gao, C.: Line art colorization based on explicit region segmentation. In: Computer Graphics Forum. vol. 40, pp. 1–10. Wiley Online Library (2021)
3. Cao, Y., Meng, X., Mok, P., Lee, T.Y., Liu, X., Li, P.: AnimeDiffusion: Anime diffusion colorization. IEEE Transactions on Visualization and Computer Graphics **30**(10), 6956–6969 (2024)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
5. Casey, E., Pérez, V., Li, Z.: The animation transformer: Visual correspondence via segment matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11323–11332 (2021)
6. Dai, Y., Li, Q., Zhou, S., Luo, Y., Li, C., Loy, C.C.: Paint bucket colorization using anime character color design sheets. arXiv preprint arXiv:2410.19424 (2024)
7. Dai, Y., Zhou, S., Li, Q., Li, C., Loy, C.C.: Learning inclusion matching for animation paint bucket colorization. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 25544–25553 (2024)
8. Feng, X., Huang, T., Wang, P., Huang, Z., Haihang, Z., Zou, Y., Li, D., Zou, K.: A unified framework for industrial cel-animation colorization with temporal-structural awareness. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19301–19310 (October 2025)
9. Google: Nano Banana 2: Combining Pro capabilities with lightning-fast speed — blog.google. `https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/`, [Accessed 21-06-2026]
10. Guajardo, J., Bursalioglu, O., Goldman, D.B.: Generative ai for 2d character animation. In: ACM SIGGRAPH 2024 Posters, pp. 1–2 (2024)
11. Huang, Z., Zhang, M., Liao, J.: LVCD: reference-based lineart video colorization with diffusion models. ACM Transactions on Graphics (TOG) **43**(6), 1–11 (2024)
12. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: CoTracker: It is better to track together. In: European Conference on Computer Vision. pp. 18–35. Springer (2024)
13. Kim, S., Park, D., Shim, B.: Semantic-aware superpixel for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1142–1150 (2023)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
15. Krause, A., Golovin, D.: Submodular function maximization. Tractability **3**(71-104), 3 (2014)
16. Li, J., Liang, Q., Li, Q., Gang, R., Fang, J., Lin, C., Feng, S., Liu, X.: RTTLC: Video colorization with restored transformer and test-time local converter. pp. 1722–1730 (06 2023). `https://doi.org/10.1109/CVPRW59228.2023.00173`
17. Li, L., Wang, G., Zhang, Z., Li, Y., Li, X., Dou, Q., Gu, J., Xue, T., Shan, Y.: Tooncomposer: Streamlining cartoon production with generative post-keyframing. arXiv preprint arXiv:2508.10881 (2025)

18. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: 2023 IEEE/CVF International Conference on Computer Vision. pp. 9264–9275. IEEE (2023)
19. Liu, Z., Cheng, K.L., Chen, X., Xiao, J., Ouyang, H., Zhu, K., Liu, Y., Shen, Y., Chen, Q., Luo, P.: Manganinja: Line art colorization with precise reference following. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5666–5677 (2025)
20. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (Nov 2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94, https://doi.org/10.1023/B:VISI.0000029664.99615.94
21. Maejima, A., Kubo, H., Funatomi, T., Yotsukura, T., Nakamura, S., Mukaigawa, Y.: Graph matching based anime colorization with multiple references. In: ACM SIGGRAPH 2019 Posters, pp. 1–2 (2019)
22. Manli, S., Weili, N., De-An, H., Zhiding, Y., Tom, G., Anima, A., Chaowei, X.: Test-time prompt tuning for zero-shot generalization in vision-language models. In: Advances in Neural Information Processing Systems (2022)
23. Meng, Y., Ouyang, H., Wang, H., Wang, Q., Wang, W., Cheng, K.L., Liu, Z., Shen, Y., Qu, H.: AniDoc: Animation creation made easier. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18187–18197 (2025)
24. Nagata, K., Kaneko, N.: DACoN: Dino for anime paint bucket colorization with any number of reference images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17899–17908 (2025)
25. Nakanishi, H., Shichijo, N., Sugi, M., Ogata, T., Hara, T., Ota, J.: Modeling the process of animation production. Int. J. Autom. Technol. **7**(4), 439–450 (2013)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
27. Qiao, R., Tan, Q., Yang, M., Dong, G., Yang, P., Lang, S., Wan, E., Wang, X., Xu, Y., Yang, L., et al.: V-thinker: Interactive thinking with images. arXiv preprint arXiv:2511.04460 (2025)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
29. Radovanovic, M., Nanopoulos, A., Ivanovic, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of machine learning research **11**(sept), 2487–2531 (2010)
30. Ramassamy, S., Kubo, H., Funatomi, T., Ishii, D., Maejima, A., Nakamura, S., Mukaigawa, Y.: Pre-and post-processes for automatic colorization using a fully convolutional network. In: SIGGRAPH Asia 2018 Posters, pp. 1–2 (2018)
31. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (June 2022)
33. Sadihin, B.C., Meng, Y., Wang, M.H., Chen, M.J., Su, H.: TimeColor: Flexible reference colorization via temporal concatenation. arXiv preprint arXiv:2601.00296 (2026)

34. Schuurmans, M., Berman, M., Blaschko, M.B.: Efficient semantic image segmentation with superpixel pooling. arXiv preprint arXiv:1806.02705 (2018)
35. Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J.: Better aggregation in test-time augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1214–1223 (2021)
36. Shi, M., Zhang, J.Q., Chen, S.Y., Gao, L., Lai, Y.K., Zhang, F.L.: Reference-based deep line art video colorization. IEEE Transactions on Visualization and Computer Graphics **29**(6), 2965–2979 (2022)
37. Shlapentokh-Rothman, M., Blume, A., Xiao, Y., Wu, Y., TV, S., Tao, H., Lee, J.Y., Torres, W., Wang, Y.X., Hoiem, D.: Region-based representations revisited. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 17107–17116 (2024)
38. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data **6**(1), 1–48 (2019)
39. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
40. Smith, A.R.: Tint fill. In: Proceedings of the 6th Annual Conference on Computer Graphics and Interactive Techniques. p. 276–283. SIGGRAPH '79, Association for Computing Machinery, New York, NY, USA (1979). https://doi.org/10.1145/800249.807456, https://doi.org/10.1145/800249.807456
41. Tang, Y., Guo, J., Liu, P., Wang, Z., Hua, H., Zhong, J.X., Xiao, Y., Huang, C., Song, L., Liang, S., et al.: Generative AI for cel-animation: A survey. arXiv preprint arXiv:2501.06250 (2025)
42. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: SIGLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
43. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=uXl3bZLkr3c
44. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 2566–2576 (2019)
45. Wittgenstein, L.: Remarks on Colour. University of California Press (1977), https://books.google.co.uk/books?id=xQhbwgEACAAJ
46. Xing, J., Liu, H., Xia, M., Zhang, Y., Wang, X., Shan, Y., Wong, T.T.: Tooncrafter: Generative cartoon interpolation. ACM Transactions on Graphics (TOG) **43**(6), 1–11 (2024)
47. Yang, Y., Fan, L., Lin, Z., Wang, F., Zhang, Z.: Layeranimate: Layer-level control for animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10865–10874 (October 2025)
48. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)
49. Zhang, Y., Ma, Y., Wang, B., Chen, Q., Wang, Z.: Follow-your-color: Multi-instance sketch colorization (2025), https://arxiv.org/abs/2503.16948

50. Zhang, Y., Wang, L., Wang, H., Wu, D., Lin, Z., Wang, F., Song, L.: Animecolor: Reference-based animation colorization with diffusion transformers. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 6682–6690 (2025)
51. Zhao, H., Cai, Z., Si, S., Ma, X., An, K., Chen, L., Liu, Z., Wang, S., Han, W., Chang, B.: MMICL: Empowering vision-language model with multi-modal in-context learning. In: The Twelfth International Conference on Learning Representations
52. Zhao, Y., Zheng, H., Luo, J., Lam, E.Y.: Improving video colorization by test-time tuning. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 166–170. IEEE (2023)
53. Zhuang, J., Ju, X., Zhang, Z., Liu, Y., Zhang, S., Yuan, C., Shan, Y.: ColorFlow: Retrieval-augmented image sequence colorization. arXiv preprint arXiv:2412.11815 (2024)
54. Zhuang, J., Li, L., Ju, X., Zhang, Z., Yuan, C., Shan, Y.: Cobra: Efficient line art colorization with broader references. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)