Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models

CVPR 2024 (Oral)

Daniel Geng Inbum Park Andrew Owens
University of Michigan

STRUCT Group Seminar
Presenter: Lan Xicheng
2024.9.14

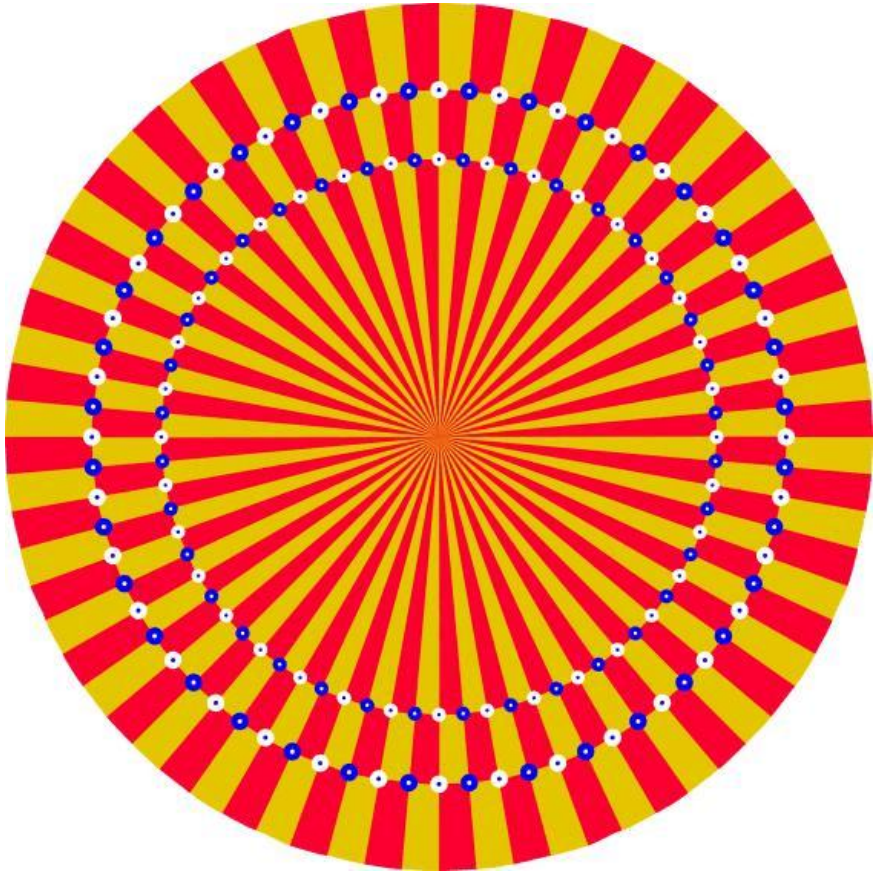
Outline

- Authors
- Background
- Methods
- Experiments
- Conclusion

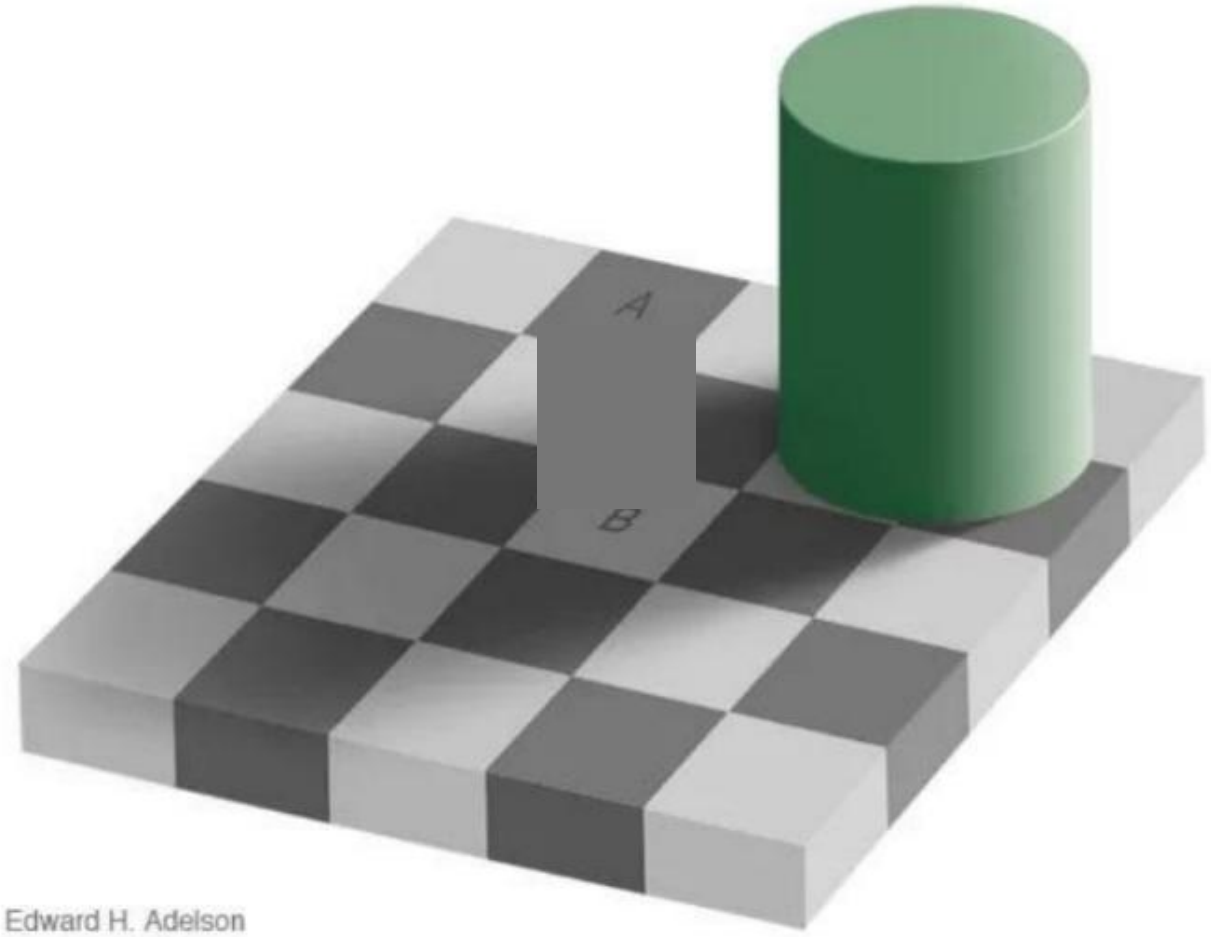
Outline

- Authors
- Background
- Methods
- Experiments
- Conclusion

Background: Optical Illusions

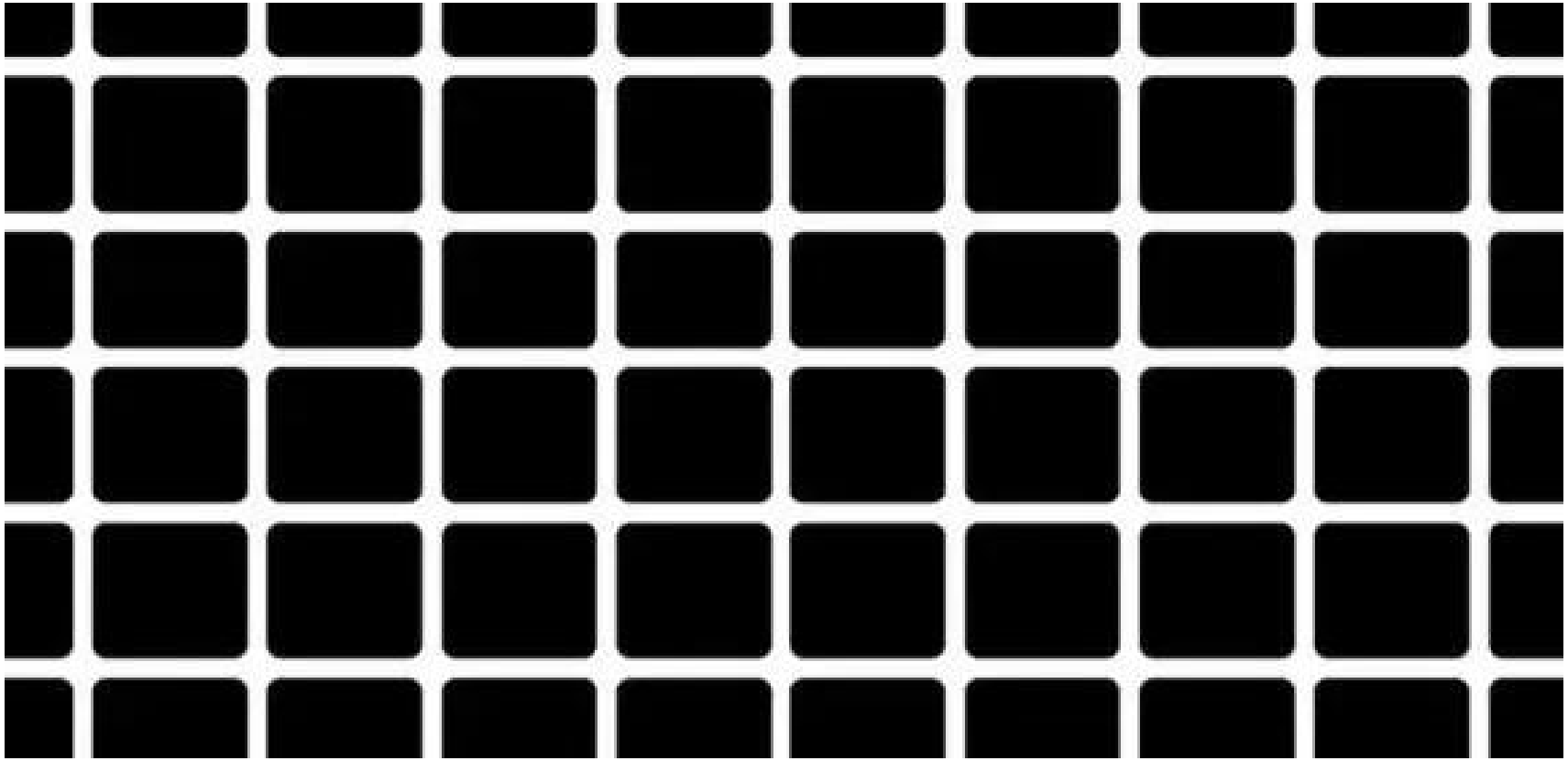


Background: Optical Illusions



Edward H. Adelson

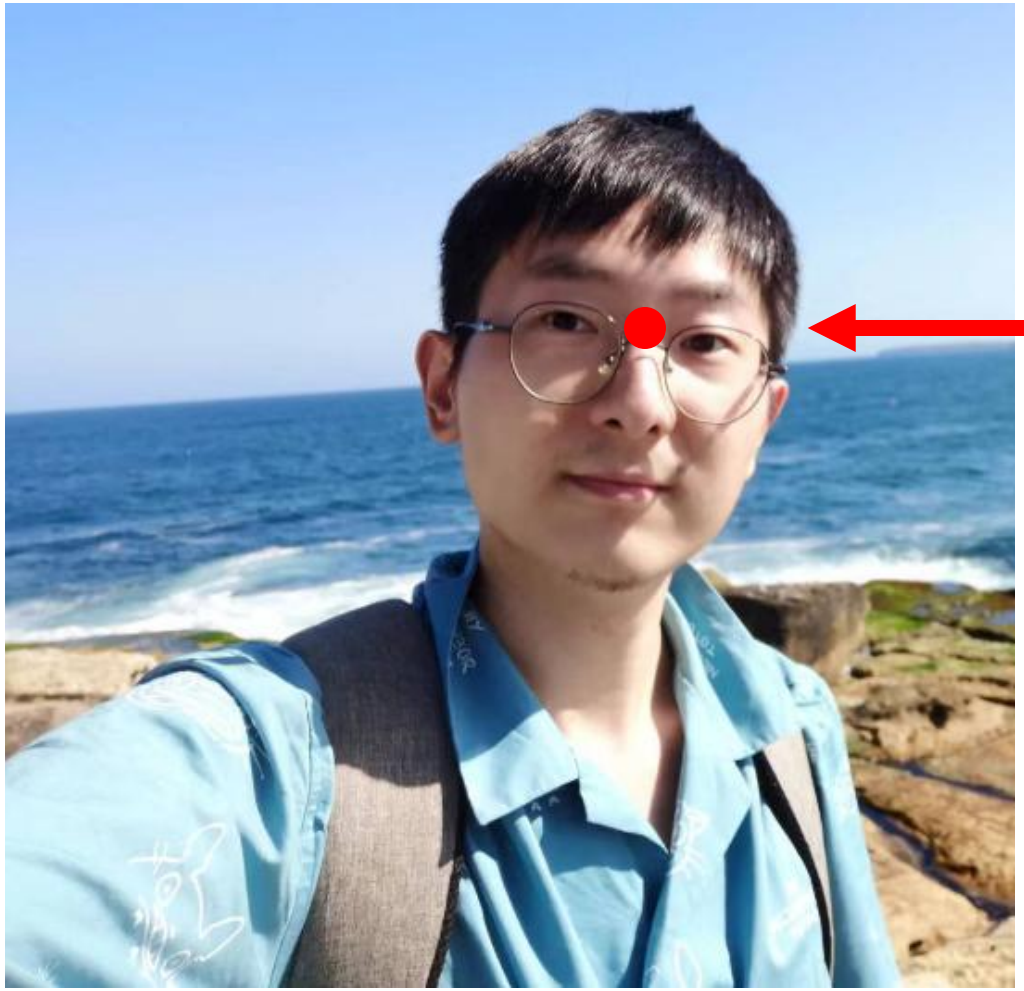
Background: Optical Illusions



Background: Optical Illusions

Stare at this point for a minute

Background: Optical Illusions

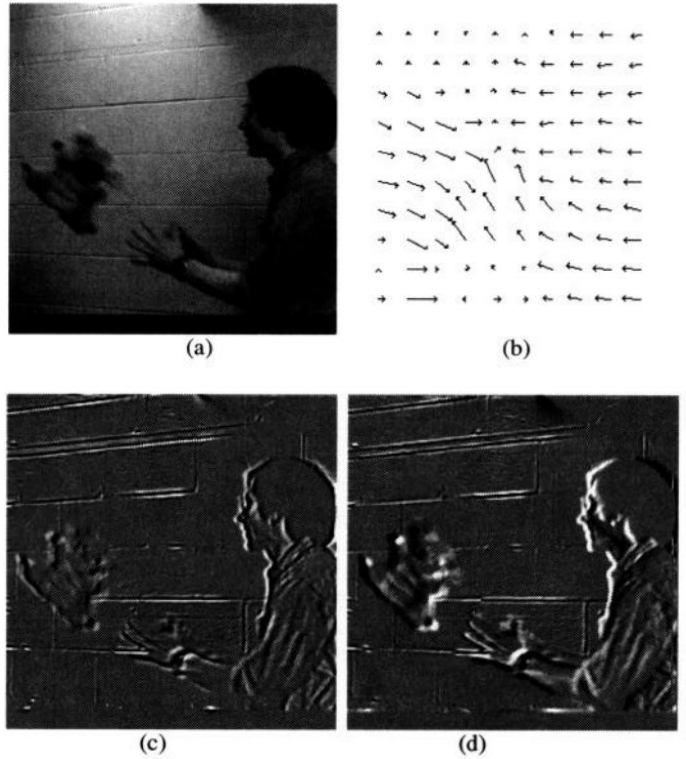


Stare at this point and blink

Background: Optical Illusions

Computational Optical Illusions

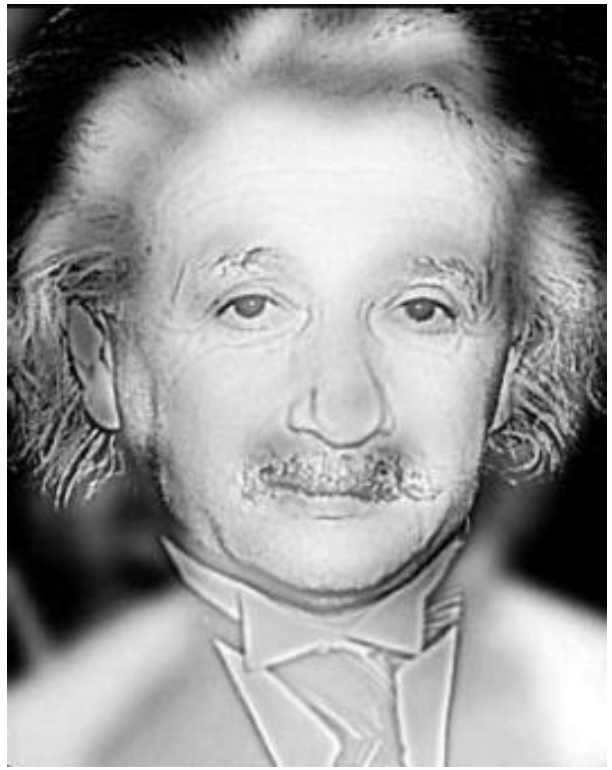
- Motion Without Movement (SIGGRAPH 1991)
  Create the illusion of constant motion in a desired direction by locally applying a filter with continuously shifting phase
- Hybrid Images (TOG 2006)
- Camouflage Images (TOG 2010)
- Designing Perceptual Puzzles By Differentiating Probabilistic Programs (SIGGRAPH 2022)



Background: Optical Illusions

Computational Optical Illusions

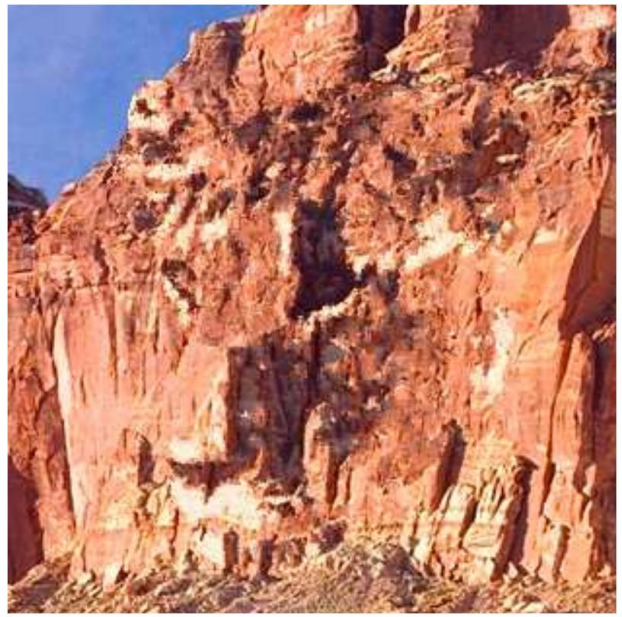
- Motion Without Movement (SIGGRAPH 1991)
- Hybrid Images (TOG 2006)
  - Change appearance depending on the distance they are viewed from
- Camouflage Images (TOG 2010)
- Designing Perceptual Puzzles By Differentiating Probabilistic Programs (SIGGRAPH 2022)



Background: Optical Illusions

Computational Optical Illusions

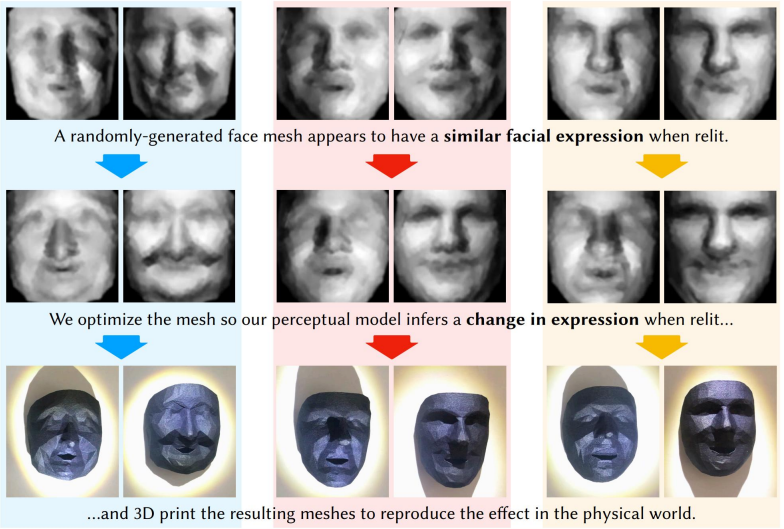
- Motion Without Movement (SIGGRAPH 1991)
- Hybrid Images (TOG 2006)
- Camouflage Images (TOG 2010)
  - Camouflage objects in a scene through retexturing, with additional constraints on luminance as to preserve salient features of the object
- Designing Perceptual Puzzles By Differentiating Probabilistic Programs (SIGGRAPH 2022)



Background: Optical Illusions

Computational Optical Illusions

- Motion Without Movement (SIGGRAPH 1991)
- Hybrid Images (TOG 2006)
- Camouflage Images (TOG 2010)
- Designing Perceptual Puzzles By Differentiating Probabilistic Programs (SIGGRAPH 2022)
    - Color-constancy, size constancy, and face perception illusions by differentiating through a Bayesian model of human vision



Background: Optical Illusions

Illusions with Diffusion Models

- QR Codes As Created Images
  - Global structure subtly matches a given template image (I2I)
- Extensions

Background: Optical Illusions

Illusions with Diffusion Models

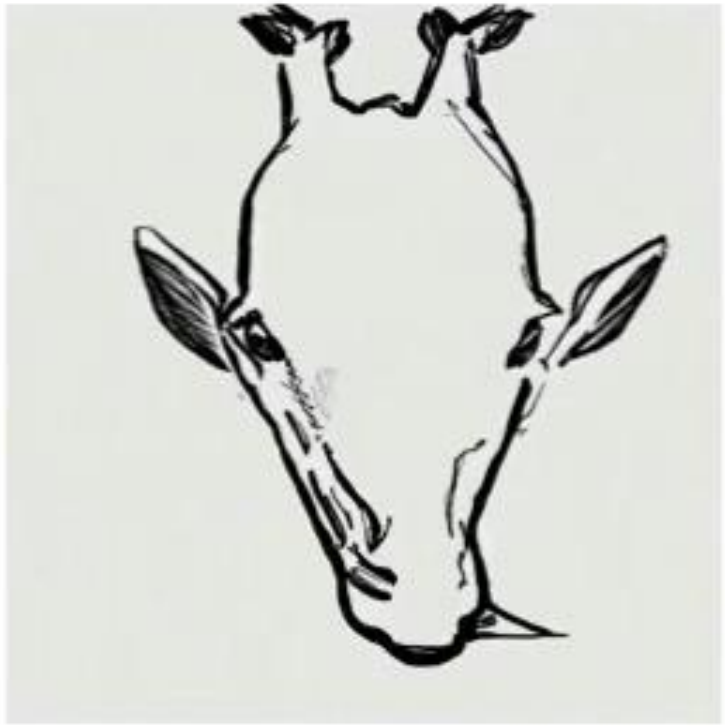
- QR Codes As Created Images
- Extensions

Background: Optical Illusions

Illusions with Diffusion Models

- QR Codes As Created Images
- Extensions

Background: Optical Illusions



a drawing of a giraffe

Background: Optical Illusions

an oil painting of
a snowy mountain village

Background: Optical Illusions

an oil painting of a skull

Background: Optical Illusions

a pop art of
marilyn monroe

Background: Optical Illusions

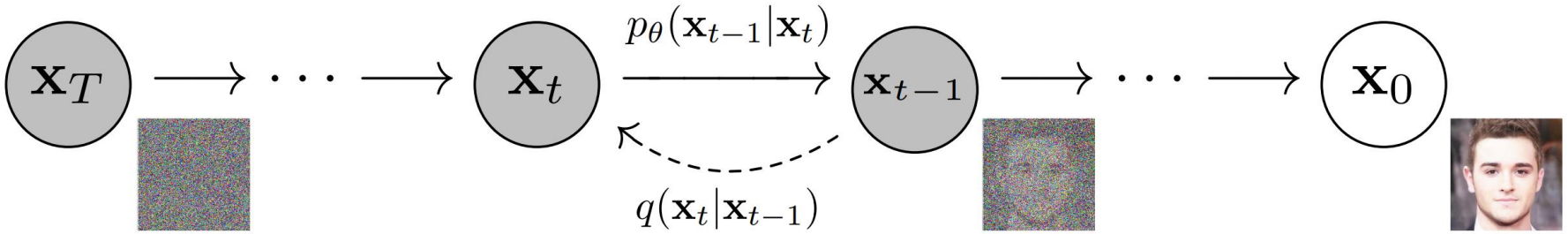
a lithograph of
a mountain landscape

Background: Optical Illusions

a watercolor of a kitten

Background

Diffusion Models



Algorithm 1 Training

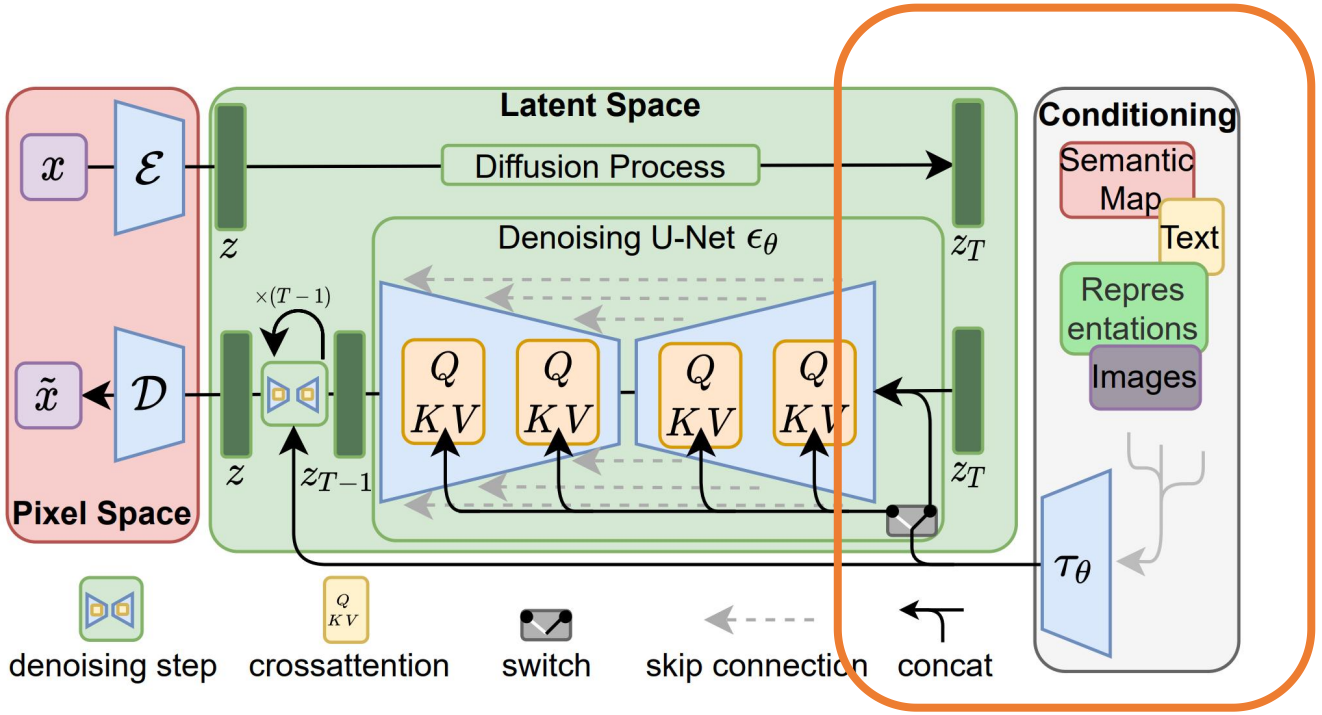
1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

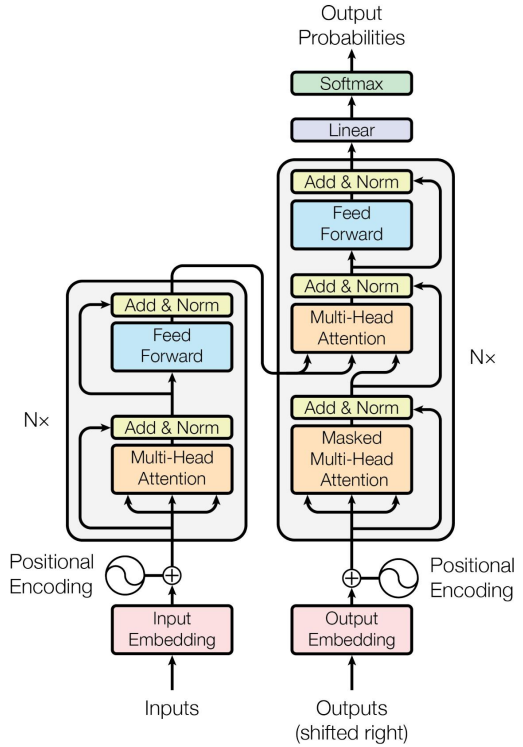
Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Background

Text-to-Image

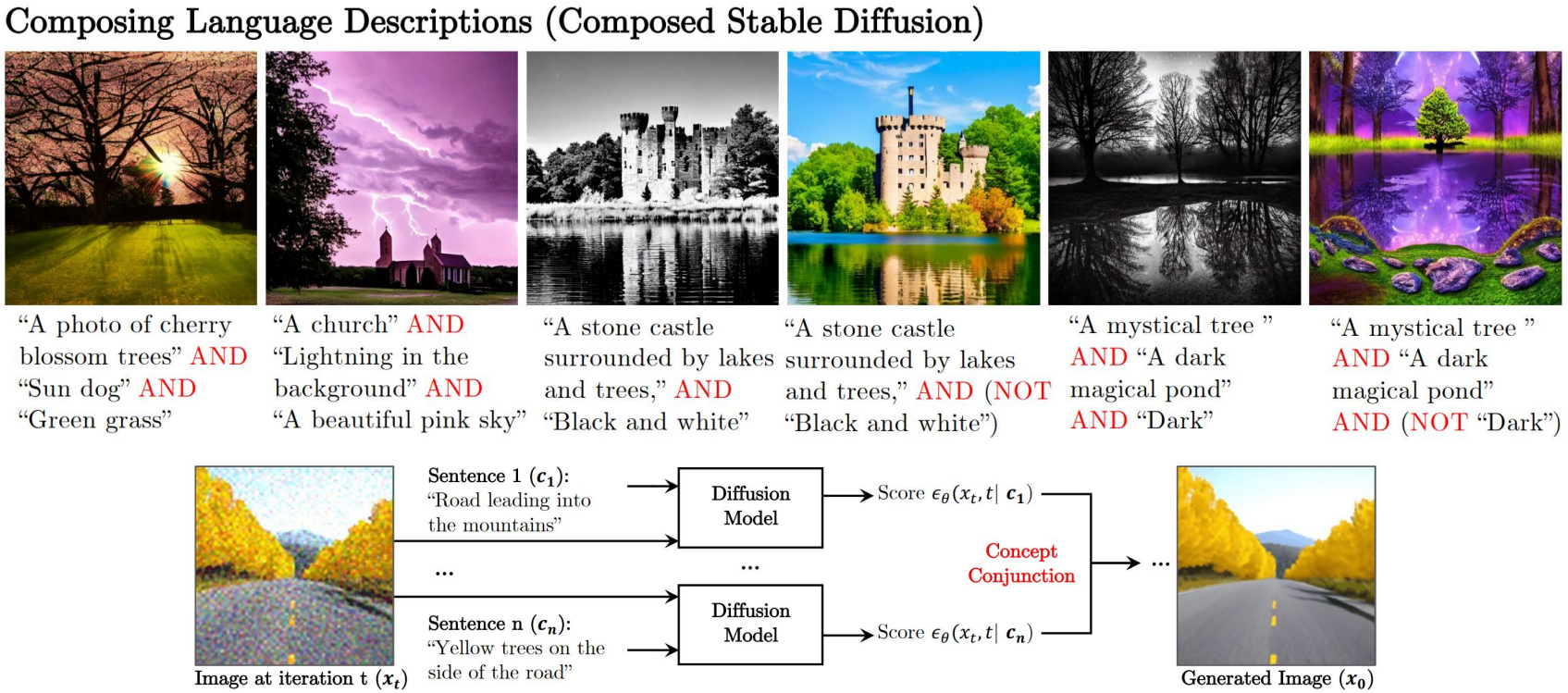




Background

Compositional Generation

- Composable Diffusion Models (ECCV 2022)
- Reduce, Reuse, Recycle (PMLR 2023)



Background

Compositional Generation

- Composable Diffusion Models (ECCV 2022)
- Reduce, Reuse, Recycle (PMLR 2023)

$$p_\theta(\boldsymbol{x}) \propto e^{-E_\theta(\boldsymbol{x})}$$

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \frac{\lambda}{2}\nabla_{\boldsymbol{x}} E_\theta(\boldsymbol{x}_{t-1}) + \mathcal{N}(0, \sigma_t^2 I)$$

Energy-Based Models (EBMs)

$$p_{\text{compose}}(\boldsymbol{x}) \propto p_\theta^1(\boldsymbol{x}) \cdots p_\theta^n(\boldsymbol{x}) \propto e^{-\sum_{i=1}^n E_\theta^i(\boldsymbol{x})}$$

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \frac{\lambda}{2}\nabla_{\boldsymbol{x}} \left(\sum_{i=1}^n E_\theta^i(\boldsymbol{x}_{t-1}) \right) + \mathcal{N}(0, \sigma_t^2 I)$$

Composing EBMs

$$p_{\text{compose}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{x}_t - \sum_{i=1}^n \epsilon_\theta^i(\boldsymbol{x}_t, t), \sigma_t^2 I \right)$$

Composing Diffusion Models

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Diffusion Models

$\boldsymbol{x}$: Image

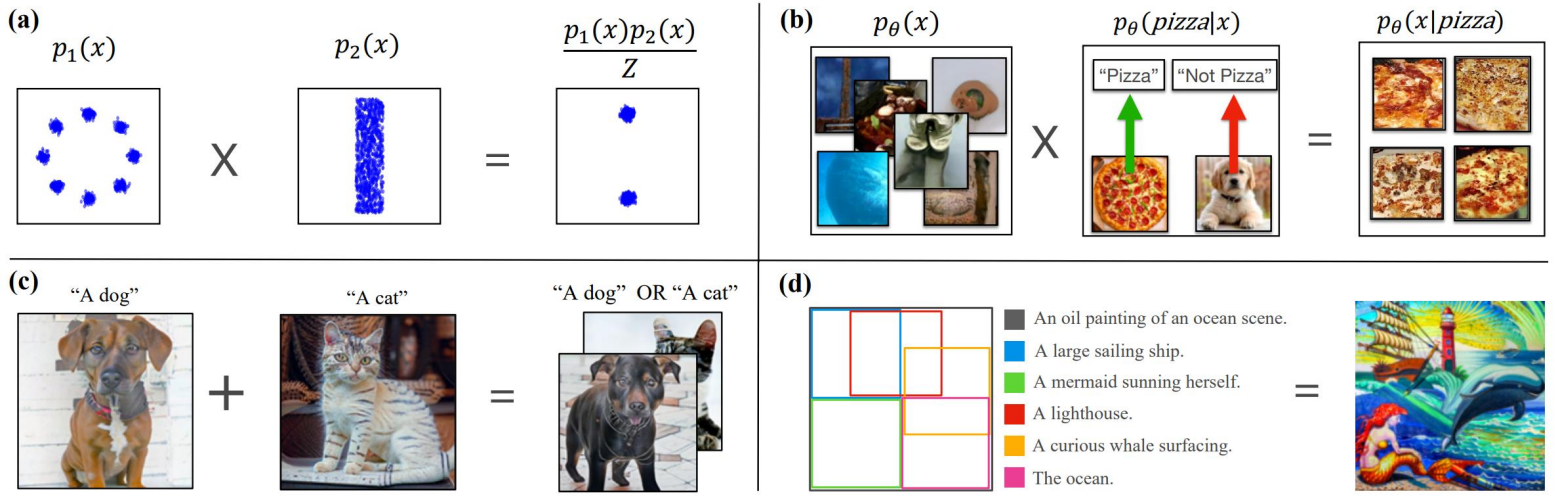
$E_\theta(\boldsymbol{x})$: Energy Function, Learnable

ϵ_θ^i : Diffusion Model

Background

Compositional Generation

- Composable Diffusion Models (ECCV 2022)
- Reduce, Reuse, Recycle (PMLR 2023)



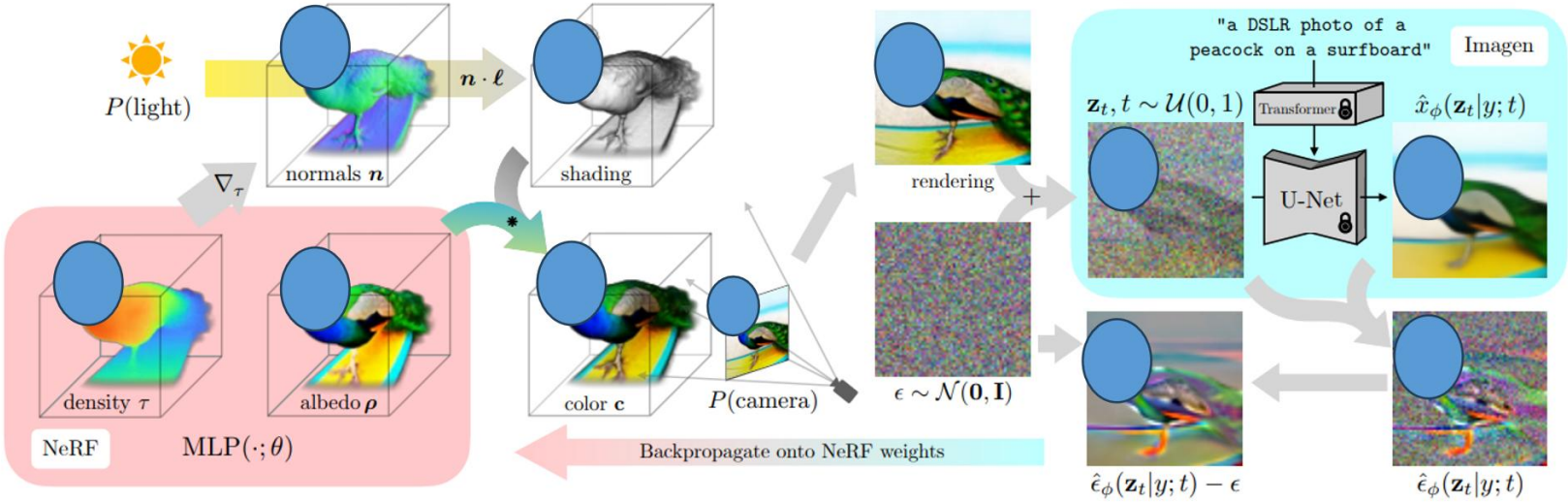
More theoretical, More in-depth, Not just relying on diffusion

Background

Illusions with Diffusion Models

Baseline: Score Distillation Sampling (SDS)

Create images align with different prompts from different views



Background

Illusions with Diffusion Models

Baseline: Matthew Tancik. Illusion diffusion. https://github.com/tancik/Illusion-Diffusion

- Combining noise predictions from different views during denoising
- Based on latent diffusion model (LDM)
- Just for rotation illusions

Outline

- Authors
- Background
- Methods
- Experiments
- Conclusion

Methods

Text-conditioned Diffusion Models

Classifier-free Guidance (CFG)

$$\epsilon_t^{\text{CFG}} = \epsilon_\theta(\mathbf{x}_t, t, \varnothing) + \gamma(\epsilon_\theta(\mathbf{x}_t, t, y) - \epsilon_\theta(\mathbf{x}_t, t, \varnothing))$$

y: Conditioning Prompts $\varnothing$: Embedding Of The Empty String γ: Strength Parameter

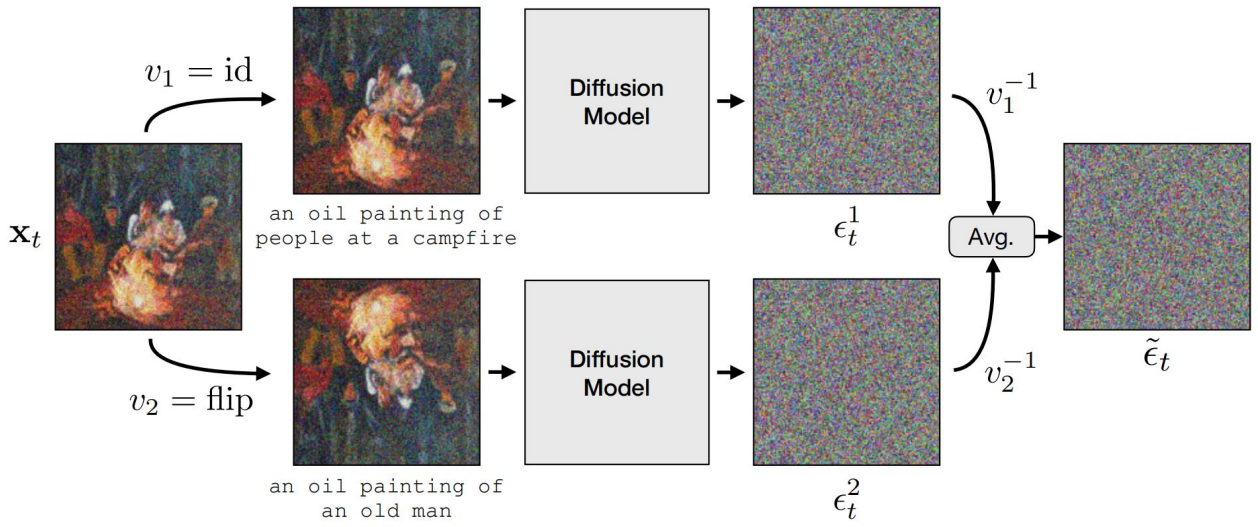
Negative prompting: the empty text prompt embedding, $\varnothing$, is replaced by a text prompt that discourage the model from generating.

Methods

Parallel Denoising

Simultaneously Denoise Multiple Views Of An Image

$$\tilde{\epsilon}_t = \frac{1}{N} \sum_i v_i^{-1} \left(\epsilon_\theta(v_i(\mathbf{x}_t), y_i, t) \right)$$



Methods

Parallel Denoising

Simultaneously Denoise Multiple Views Of An Image

$$\tilde{\epsilon}_t = \frac{1}{N} \sum_i v_i^{-1} \left(\epsilon_\theta \left(v_i(\mathbf{x}_t), y_i, t \right) \right)$$

Replace as $\epsilon_t^{\mathrm{CFG}}$

N : Number Of Prompt Set

y_i : Prompt i

$v_i(\cdot)$: View Function i

$\epsilon_\theta(\cdot)$: Diffusion Model

$$p_{\mathrm{compose}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{x}_t - \sum_{i=1}^{n} \epsilon_\theta^i(\boldsymbol{x}_t, t), \sigma_t^2 I \right)$$

Composing Diffusion Models

Methods

Conditions On Views

- Invertibility
- Linearity
- Statistical Consistency

Methods

Conditions On Views

- Invertibility
- Linearity

$$\mathbf{x}_t = w_t^{\text{signal}} \underbrace{\mathbf{x}_0}_{\text{signal}} + w_t^{\text{noise}} \underbrace{\epsilon}_{\text{noise}}$$

$v_i(\mathbf{x}_t)$ also need to be a linear combination of pure signal and pure noise with the same weighting

$$v_i(\mathbf{x}_t) = \mathbf{A}_i \mathbf{x}_t \longrightarrow \begin{aligned} v_i(\mathbf{x}_t) &= \mathbf{A}_i(w_t^{\text{signal}} \mathbf{x}_0 + w_t^{\text{noise}} \epsilon) \\ &= w_t^{\text{signal}} \underbrace{\mathbf{A}_i \mathbf{x}_0}_{\text{new signal}} + w_t^{\text{noise}} \underbrace{\mathbf{A}_i \epsilon}_{\text{new noise}} \end{aligned}$$

- Statistical Consistency

Methods

Conditions On Views

- Invertibility
- Linearity
- Statistical Consistency

-Diffusion model $\epsilon \sim \mathcal{N}(0, I)$

-Transformed noise $\mathbf{A}_i \epsilon$ must be likewise drawn from $\mathcal{N}(0, I)$

-$\mathbf{A}_i$ is an orthogonal matrix

Methods

Views Considered

Standard Image Manipulations

Rotation, reflection and skewing

an oil painting
of an old man

an oil painting of a skull

Methods

Views Considered

General Permutations

Jigsaw puzzles and inner rotations

an oil painting of a deer

an oil painting
of albert einstein

Methods

Views Considered

Color Inversion

a lithograph
of a teddy bear

a photo of a woman

Methods

Views Considered

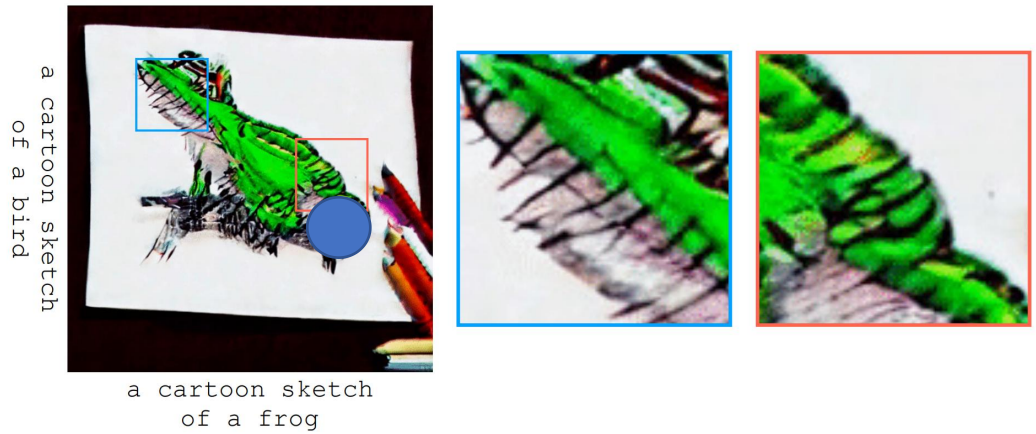
Arbitrary Orthogonal Transformations

Methods

Design Decisions

- Pixel Diffusion Model

  Latent diffusion models lead to artifacts under rotations or flips, where the location of latents change, but the content and orientation of these blocks do not



- Combining Noise Estimates
- Negative Prompting

Methods

Design Decisions

- Pixel Diffusion Model
- Combining Noise Estimates

  Alternating through them by timestep

$$\tilde{\epsilon}_t = v^{-1}_{t \bmod N}\left(\epsilon_\theta(v_{t \bmod N}(\mathbf{x}_t), t, y)\right)$$

- Negative Prompting

Methods

Design Decisions

- Pixel Diffusion Model
- Combining Noise Estimates
- Negative Prompting

  Use one view’s prompt as a negative for the other view, and vice versa
  This encourages the model to hide the other view’s prompt for a given view

Outline

- Authors
- Background
- Methods
- Experiments
- Conclusion

Experiments

Metrics

- CLIP to measure how well views align with the desired prompts
- $\mathbf{S} \in \mathbb{R}^{N \times N}$ defined as $\mathbf{S}_{ij} = \phi_{\text{img}}(v_i(\mathbf{x}))^T \phi_{\text{text}}(p_j)$

  ϕ_{img} and ϕ_{text} are the CLIP visual and textual encoders respectively

- Alignment score: $\mathcal{A} = \min \text{diag}(\mathbf{S})$, measures the worst alignment
- Concealment score: $\mathcal{C} = \frac{1}{N} \text{tr}(\text{softmax}(S/\tau))$, measures how well CLIP can classify a view

Experiments

Datasets: Prompt Pairs For 2-view Illusions

- CIFAR: 10 classes from CIFAR-10, for a total of 45 prompt pairs
- Ours: compile by hand, consists of 50 prompt pairs

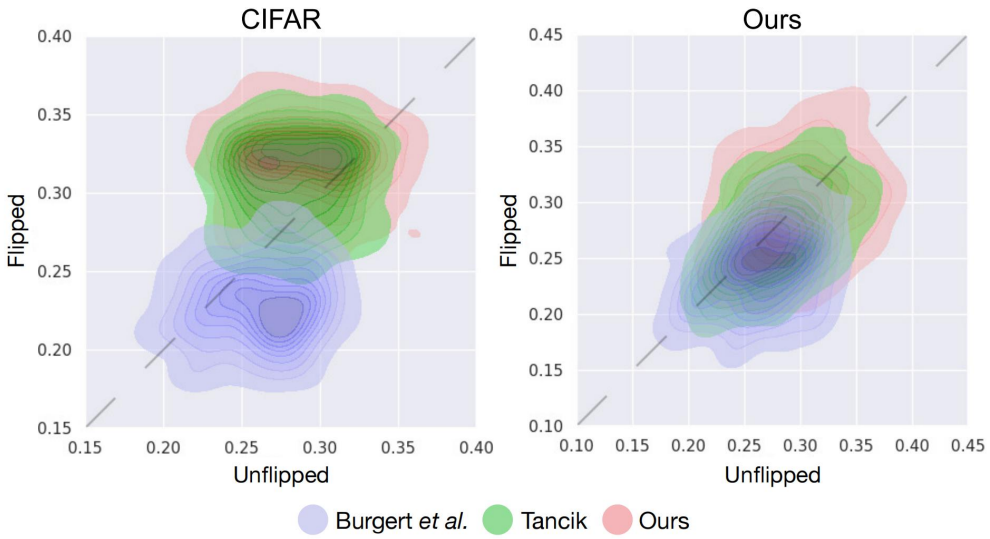
Baselines

- Burgert et al. : Score Distillation Sampling
- Tancik: An earlier version of our method

Experiments

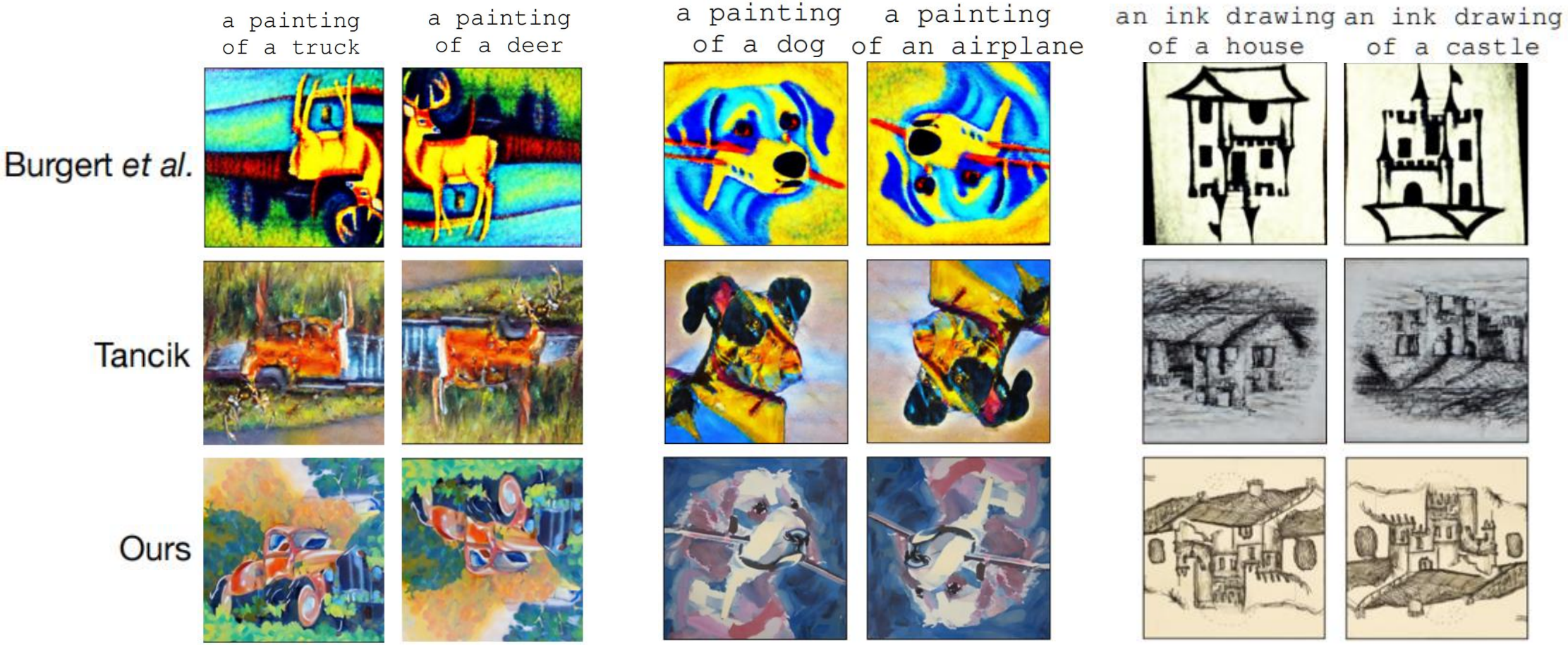
Quantitative Results

Prompt Pair	Method	$\mathcal{A}\uparrow$	$\mathcal{A}_{0.9}\uparrow$	$\mathcal{A}_{0.95}\uparrow$	$\mathcal{C}\uparrow$	$\mathcal{C}_{0.9}\uparrow$	$\mathcal{C}_{0.95}\uparrow$
CIFAR	Burgert *et al.* [2]	0.225	0.253	0.260	0.501	0.526	0.537
	Tancik [42]	0.278	0.310	0.316	0.595	0.692	0.712
	Ours	**0.287**	**0.321**	**0.327**	**0.624**	**0.717**	**0.739**
Ours	Burgert *et al.* [2]	0.233	0.270	0.283	0.501	0.526	0.538
	Tancik [42]	0.256	0.294	0.309	0.545	0.621	0.655
	Ours	**0.275**	**0.315**	**0.326**	**0.574**	**0.668**	**0.694**



Experiments

Qualitative Results



Experiments

Ablations

Ablation	$\mathcal{A} \uparrow$	$\mathcal{A}_{0.9} \uparrow$	$\mathcal{A}_{0.95} \uparrow$	$\mathcal{C} \uparrow$	$\mathcal{C}_{0.9} \uparrow$	$\mathcal{C}_{0.95} \uparrow$
Negative Prompting	0.24	0.27	0.276	**0.576**	**0.659**	**0.683**
No Negative Prompting	**0.255**	**0.285**	**0.295**	0.567	0.643	0.679
Alternating Reduction	0.252	**0.286**	0.292	0.560	0.639	0.664
Mean Reduction	**0.255**	0.285	**0.295**	**0.567**	**0.643**	**0.679**
$\gamma = 3.0$	0.239	0.271	0.285	0.537	0.610	0.629
$\gamma = 7.0$	0.255	0.285	0.295	0.567	0.643	0.679
$\gamma = 10.0$	**0.259**	**0.290**	**0.297**	**0.576**	**0.664**	**0.702**

Outline

- Authors
- Background
- Methods
- Experiments
- Conclusion

Conclusions

- Present a method to produce compelling and diverse optical illusions
- Prove the method works for a broad set of transformations
- Qualitatively show the method can generate a wide array of optical illusions

- Does not consistently produce perfect illusions

Experiments

Failures

Failure: Independent Synthesis

View: Vertical Flip

`an oil painting of a bowl of fruit`

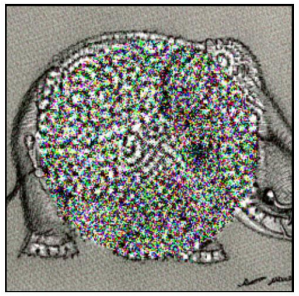
`an oil painting of a monkey`

Failure: Noise Shift

View: White Balancing

`a photo of a black and blue dress`

`a photo of a white and gold dress`



Failure: Correlated Noise

View: Rotate Inner Circle 45 Degrees (Bilinear)

`a sketch of an elephant`

`a sketch of a mouse`

Thank you for Listening!