

Learned Single-Pass Multitasking Perceptual Graphics for Immersive Displays

Doğa Yılmaz*
University College London
Department of Computer Science
London, United Kingdom
doga.yilmaz.24@ucl.ac.uk

He Wang
University College London
AI Center
Department of Computer Science
London, United Kingdom
he_wang@ucl.ac.uk

Towaki Takikawa
University of Toronto
Department of Computer Science
Toronto, Ontario, Canada
tovacinni@gmail.com

Duygu Ceylan
Adobe Research
London, United Kingdom
ceylan@adobe.com

Kaan Akşit
University College London
Department of Computer Science
London, United Kingdom
k.aksit@ucl.ac.uk

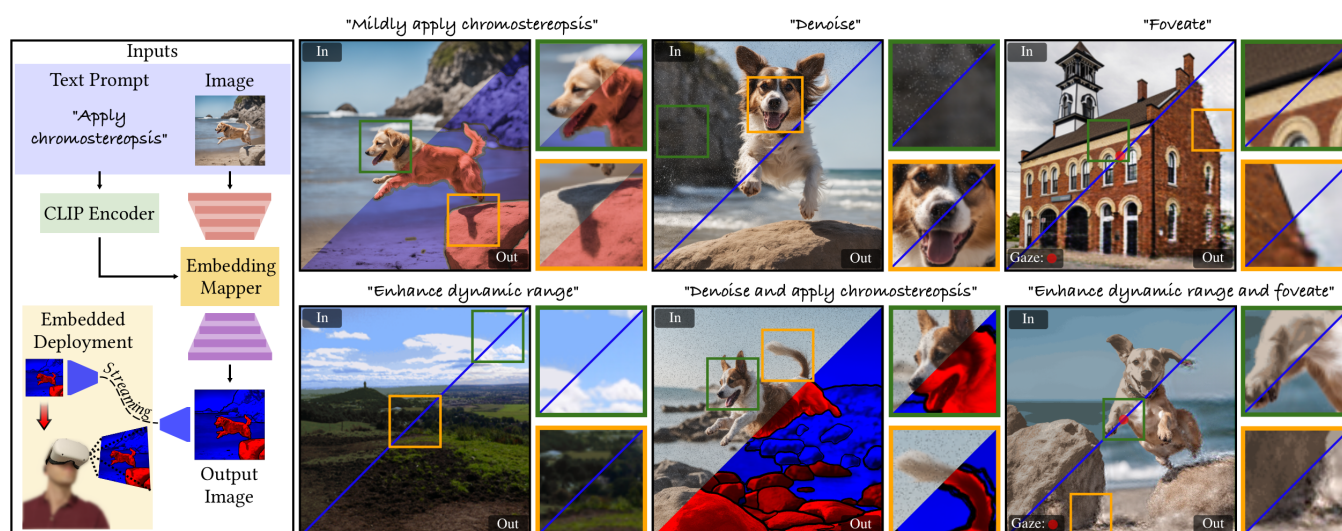


Figure 1: Our novel text-guided model (left) perceptually enhances input images based on given prompts. The model supports Chromostereopsis, Denoising, Foveation, and Dynamic Range modification, and their permutations at different intensities using adjectives like “mildly”, “slightly”, and “lightly” within a single inference, eliminating the need for daisy-chaining multiple models. Real-world input images in “foveate” and “enhance dynamic range” are attributed to Billy Wilson and Orgthingy. Remaining images are from our test set generated using Stable Diffusion [34].

Abstract

Emerging immersive display technologies efficiently utilize resources with perceptual graphics methods such as foveated rendering and denoising. Running multiple perceptual graphics methods challenges devices with limited power and computational resources. We propose a computationally-lightweight learned multitasking

perceptual graphics model. Given RGB images and text-prompts, our model performs text-described perceptual tasks in a single inference step. Simply daisy-chaining multiple models or training dedicated models can lead to model management issues and exhaust computational resources. In contrast, our flexible method unlocks consistent high quality perceptual effects with reasonable compute, supporting various permutations at varied intensities using adjectives in text prompts (e.g. “mildly”, “lightly”). Text-guidance provides ease of use for dynamic requirements such as creative processes. To train our model, we propose a dataset containing source and perceptually enhanced images with corresponding text prompts. We evaluate our model on desktop and embedded platforms and validate perceptual quality through a user study.

*Doğa Yılmaz is the corresponding author.



CCS Concepts

• **Computing methodologies** → **Perception; Image manipulation; Neural networks.**

Keywords

Perceptual Graphics, Immersive Displays, Generative Multimedia

ACM Reference Format:

Doğa Yılmaz, He Wang, Towaki Takikawa, Duygu Ceylan, and Kaan Aksit. 2025. Learned Single-Pass Multitasking Perceptual Graphics for Immersive Displays. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, October 27–31, 2025, Dublin, Ireland. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3746027.3754801>

1 Introduction

Immersive display technologies [23], including Augmented Reality (AR) glasses, Virtual Reality (VR) headsets, and large-format displays, are advancing towards more realistic image delivery.

However, these devices face constraints due to power, performance, and form-factor limitations, making on-device high-quality rendering a challenge. Hence, researchers explore perceptual graphics methods such as foveated rendering [37], dynamic range enhancement [27], image denoising [8], and chromostereopsis [40] to enhance low quality images. In practice, these emerging perceptual graphics methods need to be daisy-chained to produce images of high perceptual quality. Daisy-chaining these perceptual models or learning models for desired combined perceptual tasks can quickly lead to poor image quality as shown in Fig. 2. Moreover, these perceptual effects use the same image attributes, such as depth, segmentation, and color, which creates redundancy in the model parameters and computation. Mitigating this redundancy, a potentially more resource-efficient alternative to daisy-chaining is to combine multiple perceptual graphics methods with a multitask learning approach. Also, recent works in generative models have demonstrated that combined learned multimodal approaches, enables a wide range of image-to-image translation tasks [15, 34]. Inspired by these recent works, we propose to unify perceptual tasks in a single model to utilize bandwidth, and computational resources more efficiently while also supporting immersive displays in a device-agnostic manner, thereby meeting their unique rendering requirements.

Our work proposes a text-guided learned multitasking perceptual graphics model for immersive displays. The input to this model is an RGB image and text prompt pair to guide the model to output perceptually enhanced images. Our model is enabled by our new learned component, which we call as Embedding Mapper module. This new module efficiently combines encoded RGB images and embeddings from text prompts at the bottleneck of a

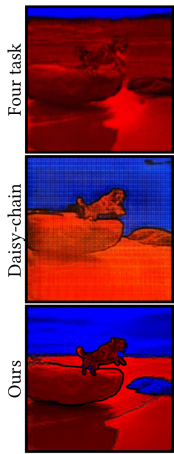


Figure 2: Four perceptual tasks applied on various baseline models and ours. The capacities are equal per task, see Sec. 4.

multitasking U-Net. Leveraging multitask learning, our model supports various perceptual tasks and their combinations, including foveated rendering, dynamic range enhancement, image denoising, and chromostereopsis. Utilizing text-embeddings rather than fixed-vectors benefits the practicality and flexibility of our model. This choice enables adjustment of effect intensity and supports new tasks without requiring any changes in the architecture. Furthermore, by maintaining a plain-text control interface, we ensure seamless compatibility with modern model-conditioning techniques. To train our model, we introduce a new dataset comprising pairs of images and their text prompts, each representing a distinct perceptual effect. Our model facilitates deployment on both desktop and embedded systems for immersive displays as it is lightweight and fast. Furthermore, we validate the perceptual quality of the images generated by our model with a subjective experiment. The source code of our learned model, along with our perceptual image dataset and model weights, can be found at https://complightlab.com/multitasking_perceptual_graphics. Our contributions are as follows:

- **Multitasking Perceptual Model.** Enabled by our new learned embedding mapper, which efficiently combines image and text embeddings, we propose a learned multitasking perceptual graphics model that transforms RGB images to various perceptually guided image styles. Our model can achieve hybrid tasks that are composed of novel permutations of individual tasks (e.g., enhance dynamic range and foveate) as well as controlling the degree of applied effect (e.g., mildly apply chromostereopsis) in a single inference step. Furthermore, we deploy our model on an NVIDIA Jetson Nano embedded device to demonstrate its effectiveness in computationally limited scenarios.
- **Perceptual Evaluations.** We introduce a new dataset that contains pairs of images and their corresponding text prompts. Each pair represents an image-to-image translation of perceptual effects. We also provide a complete pipeline describing the image generation routine in our datasets. Utilizing this dataset and image generation pipelines, we validate the perceptual quality of the generated images from our model with a user study.

2 Related Work

Our work enables the simultaneous application of multiple perceptual graphics tasks to efficiently prepare media for immersive displays by leveraging multitask learning. We review the relevant literature for each visual perception task we focus on, as well as for learned image processing methods and multitask learning approaches, to provide context for our contributions.

2.1 Visual Perception Tasks

Our work focuses on foveation, dynamic range enhancement, image denoising, and chromostereopsis. Image denoising and dynamic range enhancement are well established tasks in the literature, whereas foveation and chromostereopsis tasks are actively being explored. Here, we refer to dynamic range enhancement as increasing the bits used to represent brightness levels in an image. Following the common literature in image denoising [8, 29, 44] and dynamic range enhancement [5, 27, 43], we train our model using pairs of images with low and high dynamic range and image pairs

Table 1: Overview of perceptual graphics techniques. Our work distinguishes itself by providing support for multiple perceptual effects in a single inference pass, with text-guidance and multitasking, while maintaining a lightweight architecture. Abbreviations: Foveation (F), Chromostereopsis (C), Image Denoising (ID), Dynamic Range Enhancement (DRE).

	Approach	Perceptual Tasks	Text Guidance	Speed	Multitasking
Walton et al. [39]	Traditional	F	None	Real Time	None
Deza et al. [9]	Learned	F	None	Real Time	None
Westermann et al. [40]	Traditional	C	None	Real Time	None
Conde et al. [8]	Learned	ID	None	Real Time	None
Marzuki et al. [27]	Learned	DRE	None	Real Time	None
Affi et al. [1]	Learned	None	None	Real Time	Shared Encoder
Sun et al. [35]	Learned	None	Fixed Wording	Real Time	Hard Parameter Sharing
Huang et al. [15]	Learned	None	Open Ended	Offline	Hard Parameter Sharing
Ours	Learned	F, C, ID, DRE	Semi-Open Ended	Real Time	Hard Parameter Sharing

containing noisy and noise-free images, respectively. We provide an overview of existing perceptual graphics techniques in Tbl. 1.

Foveation. Foveated rendering promises to reduce computational complexity by rendering perceptually accurate yet lower resolution images in the periphery, leveraging the variation in resolution acuity between the fovea and periphery in the Human Visual System (HVS). Meng et al. [28] parameterizes foveated rendering by embedding polynomial kernel functions in the classic log-polar mapping, enabling variation in the sampling density and distribution of the rendered images. Another class of methods for foveated rendering uses metamers [9, 37, 39], which are image patches that are perceptually indistinguishable despite being different in terms of pixel values. Display hardware devices have recently adopted designs specifically catered towards foveated rendering [20], especially for AR and VR applications. *Our model follows the metamer approach proposed by Walton et al. [39] to foveate images.*

Chromostereopsis. Chromostereopsis [2, 30, 40] is a visual perceptual effect induced by using different colors in images, which leads to an illusion of perceived depth differences in various colors of the images. Hong et al. [14] propose an algorithm to enhance perceived depth in images based on chromostereopsis and cubic effects. Similarly, Jung et al. [17] introduce a depth map-based image enhancement algorithm utilizing chromostereopsis. Westermann et al. [40] recently proposed a novel rule-based method to enhance perceived depth in images, using results from a user study. *Building on Westermann et al. [40], our work focuses on creating artistically appealing chromostereoptic images that maximize perceived depth.*

2.2 Learned Multitasking Image Processing

Learned image-to-image translation. Isola et al. [16] investigate conditional Generative Adversarial Network (GAN) as a general-purpose solution for image-to-image translation tasks. Zhu et al. [46] propose an unpaired image-to-image translation method using a cycle-consistent approach, which mitigates the need for paired training data. Additionally, Choi et al. [6] propose a novel approach for multidomain image-to-image translations using a single model. Recently, Ko et al. [22] introduced an independent classifier to enhance feature learning, addressing the limitations of Choi et

al.’s method [6]. Ke et al. [19] propose a memory efficient learned color mapping for color normalization and stylization. Text-guided diffusion-based generative models have also been utilized for image-to-image translation [3, 12, 15, 25]. However, these generative models are iterative, requiring multiple passes to obtain good quality images, making them unsuitable for interactive speeds. For our perceptual tasks, we examined the diffusion-based approaches proposed by Brooks et al. [3] and InstructDiffusion [12]. Despite the inherent inference speed limitations, we included InstructDiffusion [12] in our evaluation as a representative state-of-the-art diffusion-based method. Additional results for Brooks et al. [3] can be found in the supplementary material (see Sec. 1). *Our work stands out as a lightweight application-specific solution suitable for embedded deployment, offering a text-guided perceptual graphics method for immersive displays.*

Multitask learning. Introduced by Caruana [4], Multitask Learning (MTL) is an inductive transfer mechanism aimed at improving generalization performance by leveraging the domain-specific information contained in the training signals of related tasks. In our work, we focus on hard parameter sharing, where all tasks share the parameters for the same model. The work by Sun et al. [36] proposes an efficient sharing scheme that learns separate execution paths for different tasks. In addition, Affi et al. [1] propose a deep multitask learning architecture for auto white balancing, utilizing a single encoder and multiple decoders, each corresponding to a specific task. Following up Affi et al. [1], Sun et al. [35] demonstrate multitasking with a single task-conditioned decoder. Alternatively, diffusion-based generative models could be utilized in multitask learning scenarios [15], but are not suitable for embedded development. *Similar to the architecture proposed by Sun et al. [35], our work utilizes text embeddings to learn multiple tasks with hard parameter sharing in the encoder and decoder. Our primary difference is in how we combine image and text embeddings using our embedding mapper.*

The aforementioned perceptual tasks and learned perceptual methods have been well explored individually. Yet, the efficient unification of these tasks into a single model remains an open challenge. *Uniquely, our solution offers a text-guided multitasking model capable of applying all these perceptual tasks within a single, fast, cohesive model that can be deployed in embedded scenarios.*

3 Text-Guided Perceptual Graphics

Given an input RGB image and a text prompt describing the desired perceptual effect, our model in Fig. 3 applies the effect such as foveation, dynamic range enhancement, image denoising, and chromostereopsis, as well as their permutations at intended scales (e.g. “mildly,” “lightly”).

Our model comprises two main components: a perceptual translation component, G , and a task-aware discriminator component, D . Firstly, our perceptual translation component, G , a modified U-Net, transforms input images into perceptually enhanced output images. This component incorporates our new Embedding Mapper module, EM , which conditions the perceptual translation based on embeddings derived from the provided input text prompts. Secondly, our discriminator, D , guides the training of the perceptual translation component, G , by verifying the outputs according to the text embeddings. During inference, D is not deployed and is utilized solely

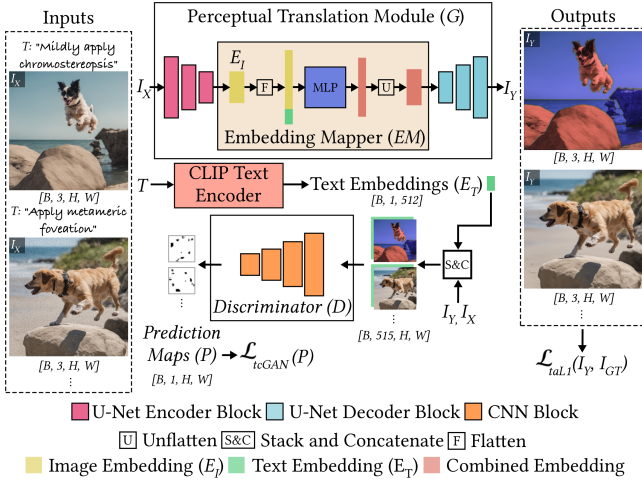


Figure 3: Our text-guided perceptual graphics model. Perceptual translation component (G) is conditioned on text embeddings (E_T) generated using CLIP [32] and Embedding Mapper (EM). The EM concatenates text embeddings (E_T) and image embeddings (E_I) to generate combined embedding (E_C). Task-aware discriminator (D) evaluates generated image (I_Y) for regularization.

during training to enhance the effectiveness of G . Additionally, for extreme cases, including low bandwidth and low power, we utilize a pre-trained autoencoder [34] to stream the generated images, in a compressed format, to the user from a server with our model.

Our model is trained using a sample size adaptive loss function to scale the loss based on the number of samples available for each task and a task-aware adversarial loss to evaluate the generated images based on the task-specific text embeddings. To train and evaluate our model, we introduce a *new* dataset comprising pairs of images, each representing a distinct perceptual enhancement.

Perceptual image translation component. Our first component, the perceptual image translation component G , has two primary objectives: (1) to enhance the visual perception of input images with desired effects, and (2) to ensure the model is lightweight and suitable for edge devices. To meet all these requirements, we employ a U-Net architecture, similar to [16], with a modified bottleneck layer, and a pre-trained CLIP model [32] for text prompt guidance. First, we transform the input text prompts, T , into text embeddings, $E_T \in \mathbb{R}^{B \times 1 \times 512}$, using the pre-trained CLIP model, CLIP. In parallel, we encode the input images, $I_X \in \mathbb{R}^{B \times 3 \times H \times W}$, into image embeddings, $E_I \in \mathbb{R}^{B \times 1 \times 2048}$, using the encoder of the U-Net. Following this, we flatten the text embeddings, E_T , and image embeddings, E_I , and concatenate them to form a single embedding, E_{T+I} . This concatenated embedding, E_{T+I} , is then fed into our embedding mapper module, EM, consisting of a Multilayer Perceptron (MLP). The EM module maps the concatenated embedding, E_{T+I} , to generate a combined embedding, $E_C \in \mathbb{R}^{B \times 1 \times 2048}$. We also derived a simple text embedding generator that maps one-hot-encoded text embeddings to the CLIP embedding space to test our model in isolation; see the supplementary Sec. 2.4 for details. This operation not only merges text and image information into a single embedding but also

ensures that the dimensionality of the resulting combined embedding, E_C , is compatible with the symmetric encoder-decoder U-Net architecture. The combined embedding, E_C , is then unflattened and fed to the bottleneck of our U-Net decoder to produce the perceptually enhanced image, $I_Y \in \mathbb{R}^{B \times 3 \times H \times W}$. Our EM offers a unique application-specific conditioning solution without requiring computationally demanding conditioning at every layer of a U-Net [18, 31] or a dedicated network for merging images and texts in the input [42] or enlarged decoder capacity due to size mismatch originated from concatenating text and image embeddings at the bottleneck [10, 38]. The architecture of our perceptual translation module, G , and our strategy for guiding the model with text prompts are illustrated in Fig. 3. Detailed configurations of EM and the pre-trained CLIP model are in the supplementary's Sec. 2.3.

Sample size adaptive loss. When introducing new tasks in the training of G , the number of samples available for a new task may be different than the existing tasks. Thus, we introduce a sample count adaptive loss function that regularizes G in training according to the number of samples available for each task. Considering the largest sample count among all tasks, SC_{MAX} , and the sample counts for each task, SC_T , we calculate boosting factors, B_T , inversely proportional to the sample counts, capped by a maximum boost coefficient, B_{MAX} . These boosting factors are then used to scale the L1 loss for each task,

$$B_T = 1 + (B_{MAX}(1 - \frac{SC_T}{SC_{MAX}})), \quad (1)$$

$$\mathcal{L}_{tal,1} = \mathcal{L}_{L1}(I_Y, I_{GT})B_T.$$

By amplifying the loss inversely proportional to the sample counts for tasks with fewer samples, we ensure optimal use of available data and encourage the optimization process to allocate greater updates for these underrepresented tasks. We further evaluate sample size adaptive L1 loss in our ablation study in Sec. 4.

Task-aware discriminator component. It has been demonstrated that utilizing conditional GAN loss [16] effectively regularizes image translation tasks by improving the quality of generated images while preserving the original content. Building on the image-based conditioning proposed by Isola et al. [16], we extend this approach to include task conditioning in our adversarial loss. To guide the training of our model, we employ a multitasking task-aware discriminator, D . This discriminator processes the generated image, I_Y , the input image, I_X , and the task-specific text embeddings, E_T , to generate probability maps that facilitates the calculation of the task-aware adversarial loss. To support this operation, the text embeddings, E_T , are stacked and concatenated along the channel dimension of the I_Y and I_X . The resulting tensor, which comprises both the image and task information, is subsequently fed into D to obtain prediction maps of $P_0 \in \mathbb{R}^{B \times 1 \times H \times W}$ and $P_1 \in \mathbb{R}^{B \times 1 \times H \times W}$. Our multitasking task-aware discriminator is illustrated in Fig. 3. We leverage P_0 and P_1 to provide a pixel-wise estimation of the likelihood that each pixel belongs to the perceptually enhanced image, I_Y . P_0 and P_1 generated by D are used to enhance task-aware guidance during the training of G as shown in the following

equation:

$$P_0 = D(I_X, I_{GT}, E_T), \quad P_1 = D(I_X, I_Y, E_T),$$

$$\mathcal{L}_{\text{tcGAN}} = \mathbb{E}_{I_X, I_{GT}, E_T} [\log P_0] + \mathbb{E}_{I_X, I_Y, E_T} [\log(1 - P_1)]. \quad (2)$$

Objective functions and training procedure. We guide the training of our model by utilizing the following functions: (1) a sample size adaptive L1 loss, and (2) a task-aware adversarial loss. Our total loss function is formulated as shown in Eq. (3), where λ is a hyperparameter that adjusts the contribution of the sample size adaptive loss to the total loss,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{tcGAN}} + \lambda \mathcal{L}_{\text{taL1}}. \quad (3)$$

We use a two-phase training strategy. Initially, our training dataset is restricted to image pairs from single tasks, which allows the model to focus on learning each task independently. After completing this phase, we expand our training dataset by incorporating both single and combined task image pairs, and continue the training process. Hyperparameters and training details are available in the supplementary’s Sec. 2.7.

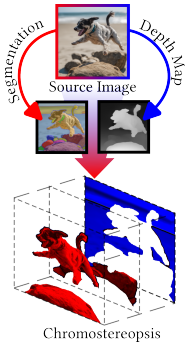


Figure 4: Utilizing depth [33] and segmentation [21] estimation to generate chromostereopsis images.

Embedding streaming component. Optionally, in scenarios where computational resources and bandwidth are extremely limited, streaming the generated images in a compressed form to the user can be beneficial for presenting high quality media. For such cases, our model can be deployed on a more powerful server, where the generated images are compressed and streamed to the client. To support this, we utilize a distilled version of the pre-trained encoder and decoder [34], where images are compressed using the encoder. The compressed images are then streamed to the client, where they are decompressed by the decoder to be displayed. This component is optional and can be disabled in scenarios where computational and bandwidth resources are not a concern. Practical details for our streaming component are available in the supplementary’s Sec. 2.5.

Perceptual graphics dataset. As our model learns a mapping from observed images, I_X , and text embeddings, E_T , to perceptually enhanced images, I_Y ,

denoted as $G : (I_X, E_T) \rightarrow I_Y$, it requires a dedicated set of text prompts paired with corresponding images for training. Generating such paired data can be particularly challenging for complex tasks like chromostereopsis, and the challenge can easily stack up as in our combined task cases. Thus, we propose a dataset containing 8800 image pairs with corresponding text prompts, each at a resolution of 1024x1024 pixels, distributed equally across various perceptual tasks, including foveated rendering, dynamic range enhancement, image denoising, chromostereopsis, and their permutations. Additionally, the prompts in our dataset feature adjectives such as “mildly,” “lightly,” and “slightly” to control the intensity of the applied effect. For all of our tasks, we generate RGB source images using Stable Diffusion [34]. Our foveated image

examples rely on Walton et al. [39], for dynamic range enhancement we clip the dynamic range of the generated images from 8 bits to 4 bits, and for image denoising we add salt and pepper noise to the generated images. To induce the chromostereopsis effect, we first generate a depth map from the ground truth images using the monocular depth estimation method by Ranftl et al. [33], then segment these images following the method by Kirillov et al. [21]. The final chromostereopsis images are produced by adjusting the hue of the foreground segments to red and the background segments to blue, based on the average depth of each segment, as proposed by Westermann et al. [40] as shown in Fig. 4. For the combined task image pairs, we apply the related methodology of each individual task consecutively. All supported tasks are listed in Tbl. 3, and details about the datasets, including the supported adjectives, can be found in the supplementary Sec. 4.

4 Evaluation

We evaluate our learned model in terms of image quality (see Tbl. 3) and inference speed (see Tbl. 2). To assess image quality, we employ metrics such as Peak Signal-to-Noise-Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity Metric (LPIPS) [45], and FovVideoVDP [26]. We conducted a user study to further confirm that the image quality produced by our model is comparable to the state-of-the-art methods. Visual results for both synthetic and real-world images are presented in Fig. 9. For more visual results beyond Fig. 9, consult our supplementary’s Sec. 10.

Inference speed. We compare our model’s performance against Sun et al. [35], InstructDiffusion [12], and relevant baseline models to assess its image quality and inference speed using 32-bit precision (FP32). Note that the method proposed by Sun et al. [35] does not support text guidance and is not fully equivalent to our proposed method in this respect. To extend the evaluation further, we formulated several baseline models, which consists of vanilla U-Net models trained on our dataset and deployed in three different settings. These settings are single-task, daisy-chain, and N-task, where N represents the number of simultaneously applied tasks. Here, a single-task model refers to a model trained for a specific task (e.g., foveation); a daisy-chain model refers to running single-task models consecutively (e.g., image denoising and foveation); and an N-task model refers to a model trained with a specific combination of tasks to perform all tasks in a single inference. For a fair comparison, we set the model capacity of all models to be equal per task (i.e. ~ 7.6 million parameters per task) as reported in Tbl. 2. Although our model has ~ 50 million parameters excluding D ; the parameters for the U-Net for our model consumes only ~ 3 million parameters, which is half the size of other baseline models with ~ 7.6 million parameters dedicated to the U-Net at minimum. The remaining ~ 47 million parameters are used for the Embedding Mapper module (EM), where the feature sizes are close to bottleneck feature size in our U-Net. *Dedicating more parameters to the EM with small feature sizes help us to achieve inference speeds as fast as a single-task model while supporting all the benefits of text-guidance and multitasking with a single model.* If the baseline models use lower capacity in their U-Net following our model (~ 3 million parameters), they render visually distorted blurry images as sampled in Fig. 5, making these

Table 2: Performance evaluation results of our model, Sun et al. [35], InstructDiffusion [12], and baselines (vanilla U-Net) on desktop (NVIDIA RTX 3090) and embedded devices (NVIDIA Jetson Nano) using 32-bit precision (FP32).

Device	Model	Inference Speed (ms)	Parameter Count (M)	Task Count
Embedded	Single-task	129.56 ms	7.656 M	1
	Daisy-chain	409.16 ms	15.312 M	2
	Daisy-chain	810.28 ms	30.624 M	4
	Two-task	129.56 ms	7.656 M	2
	Four-task	129.56 ms	7.656 M	4
	Ours (streamed)	179.14 ms	1.222 M	1-4
	Ours	260.82 ms	50.593 M	1-4
Desktop	Single-task	1.34 ms	7.656 M	1
	Daisy-chain	3.79 ms	15.312 M	2
	Daisy-chain	7.65 ms	30.624 M	4
	Two-task	1.34 ms	7.656 M	2
	Four-task	1.34 ms	7.656 M	4
	Sun et al. [35]	4.81 ms	22.124 M	1-4
	InstructDiffusion [12]	536.96 ms	859.530 M	1-4
	Ours	1.74 ms	50.593 M	1-4

lower capacity baseline models unusable for comparison. Consult our supplementary’s Sec. 2 for the lower capacity U-Net details. Our baselines (single-task, daisy-chain, and N-task) and the work by Sun et al. [35] are also limited in the number of supported tasks and do not offer full flexibility to blend tasks at will –lightly foveate and fully denoise is not an option for a foveation and denosing baseline–.

This necessitates the training of many models for many tasks, making model management an issue especially for embedded devices. In comparison to the single-task and two-task models, Tbl. 2 shows that our model has similar inference speeds. Daisy-chain models are naturally slower than our model in inference speed due to dedicating larger capacities. In contrast to our method, InstructDiffusion [12] has significantly slower inference due to its larger model size and iterative diffusion process. As observed in Tbl. 2, among flexible and controllable models, ours is the fastest, achieving inference speeds 2.5 times faster than the nearest comparable approach.

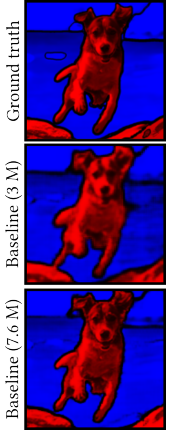


Figure 5: Lower capacity per task leads to visual distortions in our baseline models.

Tbl. 3 are averaged over the test split. Across a variety of metrics, our models achieve on-par performance in terms of image quality when compared to competitor models. Beyond image quality, our approach offers notable advantages in versatility, task adaptability, precise control over effect intensity, and the flexibility to deploy either fully or partially on an embedded device. By using

Quantitative image quality. We evaluate the visual quality of our model, method proposed by Sun et al. [35], InstructDiffusion [12], and the baselines with established metrics in Tbl. 3 and provide Fig. 9 to sample their task performance qualitatively for both synthetic and real-world images. Results in

Table 3: Quantitative image quality of our model, Sun et al. [35], InstructDiffusion [12], and baseline models (vanilla U-Nets). Abbreviations: Foveation (F), Chromostereopsis (C), Image Denoising (ID), Dynamic Range Enhancement (DRE).

Task	Model	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	FovVideoVDP ↑
F	Single-task	27.43	0.79	0.18	9.23
	Sun et al. [35]	27.12	0.78	0.23	9.17
	InstructDiffusion [12]	25.74	0.75	0.14	8.92
	Ours	25.64 (-1.79)	0.74 (-0.05)	0.10 (+0.04)	9.01 (-0.22)
DRE	Single-task	33.38	0.92	0.05	9.25
	Sun et al. [35]	31.92	0.91	0.06	9.21
	InstructDiffusion [12]	29.89	0.90	0.03	9.19
	Ours	31.07 (-2.31)	0.88 (-0.04)	0.08 (-0.03)	9.06 (-0.19)
ID	Single-task	35.90	0.95	0.03	9.79
	Sun et al. [35]	34.26	0.94	0.04	9.76
	InstructDiffusion [12]	29.49	0.83	0.07	9.53
	Ours	34.05 (-1.85)	0.92 (-0.03)	0.08 (-0.05)	9.77 (-0.02)
C	Single-task	16.87	0.81	0.14	5.53
	Sun et al. [35]	17.02	0.80	0.14	5.80
	InstructDiffusion [12]	17.03	0.75	0.13	5.92
	Ours	17.04 (+0.01)	0.81 (0.00)	0.13 (0.00)	5.54 (-0.38)
ID and C	Two-task	16.94	0.81	0.15	5.45
	Daisy-chain	16.02	0.73	0.16	5.43
	Sun et al. [35]	17.80	0.81	0.11	5.44
	InstructDiffusion [12]	17.43	0.67	0.12	4.44
DRE and C	Two-task	16.74 (-1.06)	0.80 (-0.01)	0.14 (-0.03)	5.47 (+0.02)
	Daisy-chain	16.49	0.80	0.14	5.29
	Sun et al. [35]	16.27	0.80	0.14	5.27
	InstructDiffusion [12]	16.53	0.80	0.15	5.33
ID and F	Two-task	15.91 (-1.70)	0.78 (-0.02)	0.16 (-0.04)	5.36 (+0.03)
	Daisy-chain	27.15	0.78	0.22	9.16
	Sun et al. [35]	27.15	0.78	0.20	9.17
	InstructDiffusion [12]	26.98	0.77	0.23	9.14
DRE and F	Two-task	25.65 (-1.50)	0.71 (-0.07)	0.11 (+0.08)	8.98 (-0.19)
	Daisy-chain	26.60	0.75	0.23	8.85
	Sun et al. [35]	26.59	0.76	0.21	8.84
	InstructDiffusion [12]	26.35	0.74	0.25	8.77
DRE and ID and F and C	Two-task	25.06 (-1.54)	0.69 (-0.07)	0.11 (+0.01)	8.58 (-0.27)
	Daisy-chain	24.65	0.73	0.12	8.58
	Sun et al. [35]	17.05	0.62	0.18	5.46
	InstructDiffusion [12]	17.02	0.57	0.14	5.53
	Ours	17.14 (+0.09)	0.66 (+0.04)	0.14 (+0.04)	5.61 (+0.08)

text-guidance adjectives (e.g., “strongly foveate”) to specify different effect intensities, our model can dynamically adjust perceptual effects, as demonstrated in Fig. 9. In contrast, while baseline models lack on-the-fly adaptability, Sun et al. [35] offer similar functionality with inference speed 2.5 times slower. InstructDiffusion [12] offers on-the-fly adaptability; however, its high computational demands make it unsuitable for real-time applications and deployment on embedded devices. Additional visual results for other supported tasks are provided in our supplementary’s Sec. 10. Video results of our method are also included, with further details available in Sec. 2.8.

Supporting complex tasks. The increasing number of tasks in a combined task compounds the overall complexity, making it more challenging to produce high-quality images. As indicated in the last row of Tbl. 3 and in Fig. 2, when the task count increases to four, our model surpasses both the daisy-chain method and the four-task models across all measured metrics. This indicates that the ability of the four-task model becomes insufficient for generating high-quality images. In contrast, our multitasking approach provides a flexible solution, allowing for the blending of tasks and the generation of high-quality images using a single model.

Ablation study. We conducted ablation studies to validate the effectiveness of our proposed components. Specifically, we evaluated the performance contribution of each component of our loss

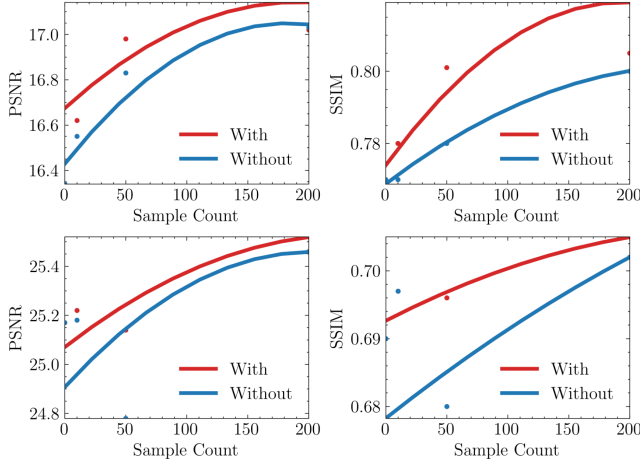


Figure 6: Performance of our model in terms of PSNR and SSIM across various training sample sizes for two reduced tasks: image denoising and chromostereopsis (top row) and denoise and foveate (bottom row). Red curves represent results with sample size adaptive loss, while blue curves represent results without it. The plotted curves are second-degree polynomials fitted to the data.

function. Additionally, we inspected how the parameter count allocated to the *EM* impacts the model’s performance. **Task-aware adversarial loss.** Fig. 7 presents a representative sample image that demonstrates the impact of the task-aware adversarial loss. When we include the task-aware adversarial loss in our model, we observe that the model preserves high-frequency details of the image.

Without the task-aware adversarial loss, the model fails to preserve these details similar to the baseline models. We invite readers to observe the high-frequency details in the foveated regions of the images in Fig. 7. **Sample size adaptive loss.** In our training dataset, we reduced the sample count of a set of tasks to simulate a low sample count scenario. The reduced sample sizes are as follows: 0, 50, 100, 150, and 200, whereas a non-reduced task contain 880 samples. We measured the performance using PSNR and SSIM (Fig. 6), showing that the sample size adaptive loss improves performance on low-sample tasks.

EM parameter count. Beginning with the smallest possible *EM* (~ 50 M parameters), constrained by image dimensions, we incrementally increased its capacity to ~ 60 M and ~ 100 M. Our observations indicate that increasing the capacity of *EM* does not affect the model’s performance. Sample images generated using *EM* of different sizes, along with their corresponding performance metrics, can be found in the supplementary’s Sec. 6.

Subjective evaluation. We conducted an informal subjective study with 22 participants (age 18–30; 5 female, 17 male), all naïve to the



Figure 7: Our task-aware adversarial loss preserve features at the periphery in foveated rendering.

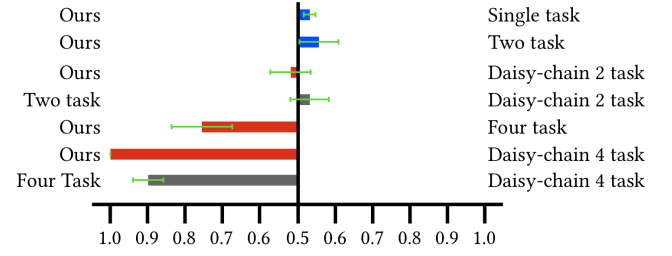


Figure 8: Preferences of participants in the user study. Our model is on-par with single-task models, two-task models, and daisy-chaining of two models, while outperforming the daisy-chaining of four tasks and the four-task model in terms of user preference. The colored bars indicate preference percentage, while green lines indicate a 95% confidence interval.

study’s purpose. The study comprised five sections, each assessing a different task using 15 image pairs. Participant preferences between our model and the baselines are summarized in Fig. 8, with 95% confidence intervals. Before each section, participants were informed about the task and asked to rate the image pairs based on their preferences. Participants’ preferences indicate that our model performs on par with single-task models, two-task models, and the daisy-chaining of two models, with half of the participants preferring our model. For the four task cases, participants preferred our model with probabilities of 100% and 74% compared to the daisy-chaining of four tasks and the four-task model, respectively. The participants’ preferences are consistent with our quantitative results in Tbl. 3 and further support our findings about complex tasks in Sec. 4. Additional details about the user study are available in the supplementary’s Sec. 5.

5 Discussion

There are various limitations and potential future research directions that may help overcome the limitations in our learned multitasking perceptual graphics model.

Visual artifacts. In a small set of test cases where adjectives such as “strongly,” “lightly,” or “mildly” are used, we observe color deviations and a minor noise from the ground truth images. In scenarios involving multiple tasks, ambiguous cases may occur when two effects target the same region of the image. Task prioritization is crucial in such cases to avoid visual artifacts. Our experiments indicate that our model tends to prioritize the task with the higher loss value, see supplementary’s Sec. 7. Among the tasks we support, chromostereopsis induces the largest change in pixel values and, as a result, is prioritized over other tasks. For the extended discussion on visual artifacts, consult our supplementary’s Sec. 9.

Task-specific visual quality metrics. Generic image quality metrics such as PSNR, SSIM, LPIPS and FovVideoVDP are not well suited for chromostereopsis and foveation cases, as shown in the first and third rows of Fig. 9. In the case of foveation, these metrics fail to capture the metameric patterns in the peripheral regions, which are essential for the task. In the case of chromostereopsis, they do not detect some artifacts that are present. This limitation restricts the ability to further improve the quality of the generated

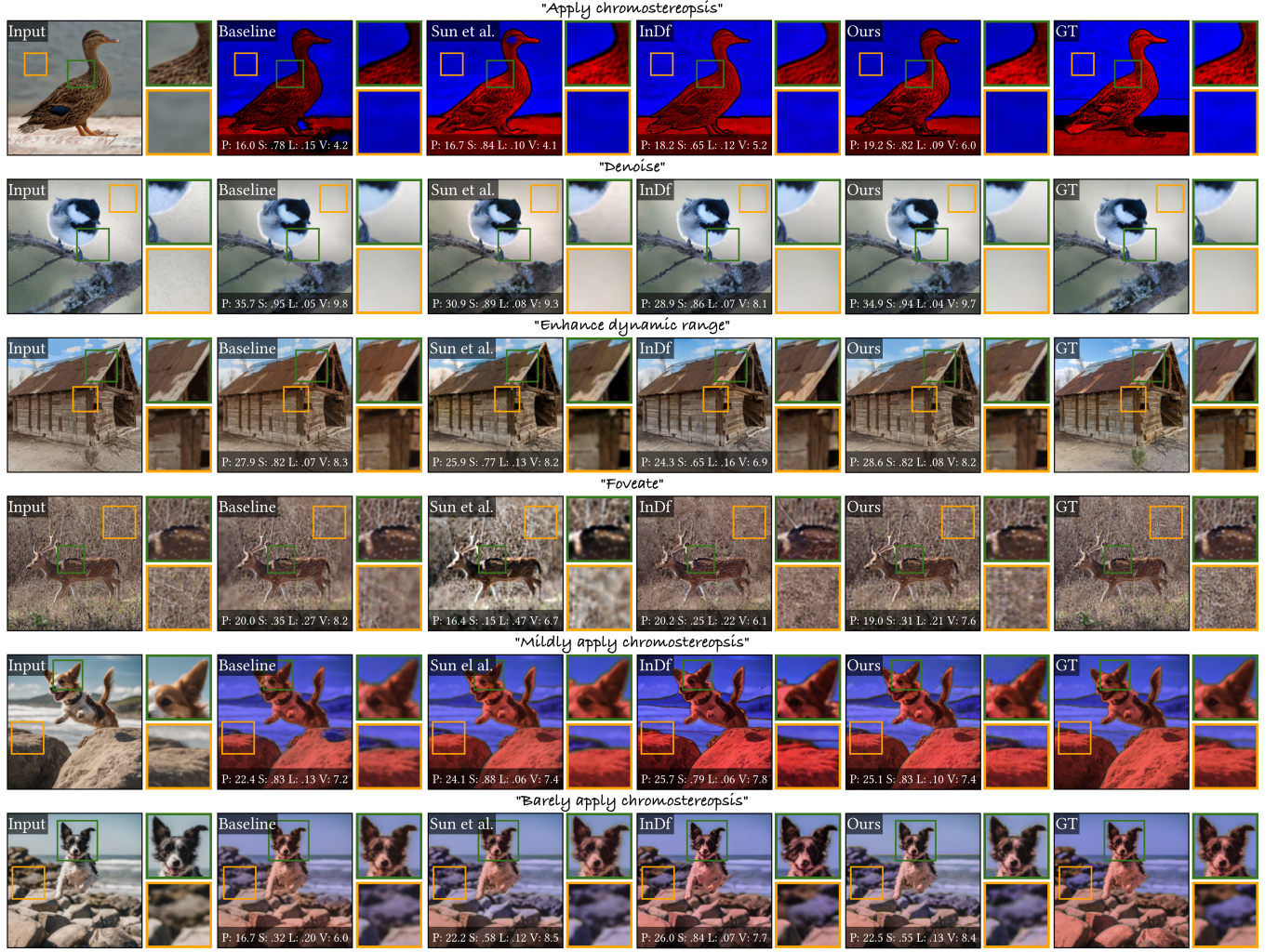


Figure 9: Results of our multitasking perceptual model compared to Sun et al. [35], InstructDiffusion (InDf) [12], and task-specific Vanilla U-Nets (Baseline). Metrics: PSNR (P) ↑ SSIM (S) ↑ LPIPS (L) ↓ FovVideoVDP (V) ↑. Dynamic range enhancement results are best seen on video displays. (Real-world images in rows 1-4 attributed to Miguel Discart, Michael Kuhn, James Marvin Phelps, and Lahar Jadav. Remaining images from our test set generated using Stable Diffusion [34].)

images using learned methods, such as super-resolution [24], as demonstrated in our supplementary Sec. 2.6.

Supporting more tasks. Our model can potentially serve as a post-processor for many real-world applications, such as video streaming, image enhancement, and immersive display technologies. For instance, several color-based perceptual tasks could be incorporated to reduce power usage [11], provide stereo view synthesis from a single image, and undertake prescription correction [13], anaglyph rendering [41], and accommodative rendering [7] in immersive displays. An extended discussion on the tasks currently supported can be found in the supplementary material’s Sec. 3.

Generalizing to unseen prompts and images. Unlike other methods [35], which utilize fixed one-hot-encoded text embeddings, our model can generalize to unseen prompts describing the supported

tasks, offering a more flexible and user-friendly operation. Examples demonstrating our model’s ability to generalize to unseen real-world images are provided in Fig. 9, and the supplementary’s Sec. 8. Examples of prompt generalization are also available in the supplementary’s Sec. 8. From these experiments, we can conclude that our model can generalize to unseen images and prompts that describe the supported tasks using novel wording. When we use a prompt that describes an unsupported task, the resulting output image shows negligible changes, validating the language conditioning capability.

Conclusion. The key finding from our model is effectively encoding images and text prompts for various perceptual tasks via multitask learning without exhausting computational resources. With the help of this key finding, our model efficiently enhances images for immersive displays including VR headsets and AR glasses.

Acknowledgements

The authors wish to thank Furkan Kınlı, Yicheng Zhan, Josef Spjut, and Morgan McGuire for fruitful discussions related to streaming and text-guidance aspects. The authors thank anonymous reviewers for their feedback.

References

- [1] Mahmoud Afifi and Michael S Brown. 2020. Deep white-balance editing. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1397–1406.
- [2] Yuejin Bai, Yana Zhang, and Zhaohui Li. 2016. Perceived depth modeling based on chromostereopsis. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 723–727.
- [3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.
- [4] Rich Caruana. 1997. Multitask learning. *Machine learning* 28 (1997), 41–75.
- [5] Jinyoung Choi and Bohyung Han. 2020. Task-aware quantization network for jpeg image compression. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 309–324.
- [6] Yunje Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [7] Steven A Cholewiak, Gordon D Love, Pratul P Srinivasan, Ren Ng, and Martin S Banks. 2017. Chromablr: Rendering chromatic eye aberration improves accommodation and realism. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–12.
- [8] Marcos V Conde, Florin Vasluianu, Javier Vazquez-Corral, and Radu Timofte. 2023. Perceptual image enhancement for smartphone real-time applications. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1848–1858.
- [9] Arturo Deza, Aditya Jonnalagadda, and Miguel Eckstein. 2017. Towards metamerism via foveated style transfer. *arXiv preprint arXiv 1705.10041* (2017).
- [10] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. 2017. Semantic image synthesis via adversarial learning. In *Proc. of the IEEE International Conference on Computer Vision*. 5706–5714.
- [11] Budmonde Duinkharjav, Kenneth Chen, Abhishek Tyagi, Jiayi He, Yuhao Zhu, and Qi Sun. 2022. Color-perception-guided display power reduction for virtual reality. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–16.
- [12] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, and Baining Guo. 2023. Instruct-Diffusion: A Generalist Modeling Interface for Vision Tasks. *CoRR abs/2309.03895* (2023). doi:10.48550/arXiv.2309.03895
- [13] Ahmet H Güzel, Jeanne Beyazian, Praneeth Chakravarthula, and Kaan Aksit. 2023. ChromaCorrect: prescription correction in virtual reality headsets through perceptual guidance. *Biomedical Optics Express* 14, 5 (2023), 2166–2180.
- [14] Ji Young Hong, Ho Young Lee, Du Sik Park, and Chang Yeong Kim. 2011. Depth perception enhancement based on chromostereopsis. In *Human Vision and Electronic Imaging XVI*, Vol. 7865. SPIE, 367–376.
- [15] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv 2302.09778* (2023).
- [16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proc. of the IEEE conference on Computer Vision and Pattern Recognition*. 1125–1134.
- [17] Seung-Won Jung and Sung-Jea Ko. 2012. Depth map based image enhancement using color stereopsis. *IEEE Signal Processing Letters* 19, 5 (2012), 303–306.
- [18] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proc. of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 4401–4410.
- [19] Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson WH Lau. 2023. Neural preset for color style transfer. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14173–14182.
- [20] Jonghyun Kim, Youngmo Jeong, Michael Stengel, Kaan Aksit, Rachel A Albert, Ben Boudaoud, Trey Greer, Joohwan Kim, Ward Lopes, Zander Majercik, et al. 2019. Foveated AR: dynamically-foveated augmented reality display. *ACM Trans. Graph.* 38, 4 (2019), 99–1.
- [21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv 2304.02643* (2023).
- [22] Kanghyeok Ko, Taesun Yeom, and Minhyeok Lee. 2023. Superstargan: Generative adversarial networks for image-to-image translation in large-scale domains. *Neural Networks* 162 (2023), 330–339.
- [23] George Alex Koulrieris, Kaan Aksit, Michael Stengel, Rafal K Mantiuk, Katerina Mania, and Christian Richardt. 2019. Near-eye display and tracking technologies for virtual and augmented reality. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 493–519.
- [24] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *Proc. of the IEEE conference on Computer Vision and Pattern Recognition workshops*. 136–144.
- [25] Weifeng Lin, Xinyu Wei, Renrui Zhang, Le Zhuo, Shitian Zhao, Siyuan Huang, Huan Teng, Junlin Xie, Yu Qiao, Peng Gao, et al. 2025. Pixwizard: Versatile image-to-image visual assistant with open-language instructions. *Proc. of the International Conference on Learning Representations* (2025).
- [26] Rafal K Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. 2021. Fovvideovp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–19.
- [27] Ismail Marzuki and Donggyu Sim. 2020. Perceptual adaptive quantization parameter selection using deep convolutional features for HEVC encoder. *IEEE Access* 8 (2020), 37052–37065.
- [28] Xiaoxu Meng, Ruofei Du, Matthias Zwicker, and Amitabh Varshney. 2018. Kernel Foveated Rendering. *Proc. ACM Comput. Graph. Interact. Tech.* 1, 1, Article 5 (jul 2018), 20 pages. doi:10.1145/3203199
- [29] Guy Ohayon, Theo Adrai, Gregory Vaksman, Michael Elad, and Peyman Milanfar. 2021. High perceptual quality image denoising with a posterior sampling cgan. In *Proc. of the IEEE/CVF International Conference on Computer Vision*. 1805–1813.
- [30] Maris Ozolinsh and Kristine Muizniece. 2015. Color Difference Threshold of Chromostereopsis Induced by Flat Display Emission. *Frontiers in Psychology* 6 (2015). doi:10.3389/fpsyg.2015.00337
- [31] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proc. of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [33] René Ranftl, Alexey Bochkovskiy, and Vladen Koltun. 2021. Vision transformers for dense prediction. In *Proc. of the IEEE/CVF International Conference on Computer Vision*. 12179–12188.
- [34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 10684–10695.
- [35] Guolei Sun, Thomas Probst, Danda Pani Paudel, Nikola Popović, Menelaos Kanakis, Jagruti Patel, Dengxin Dai, and Luc Van Gool. 2021. Task switching network for multi-task learning. In *Proc. of the IEEE/CVF International Conference on Computer Vision*. 8291–8300.
- [36] Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko. 2020. Adashare: Learning what to share for efficient deep multi-task learning. *Advances in Neural Information Processing Systems* 33 (2020), 8728–8740.
- [37] Taimoor Tariq and Piotr Didyk. 2024. Towards Motion Metamers for Foveated Rendering. *ACM Trans. Graph.* (2024).
- [38] Duc Minh Vo and Akihiro Sugimoto. 2022. Paired-D++ GAN for image manipulation with text. *Machine Vision and Applications* 33, 3 (2022), 45.
- [39] David R Walton, Rafael Kuffner Dos Anjos, Sebastian Friston, David Swapp, Kaan Aksit, Anthony Steed, and Tobias Ritschel. 2021. Beyond blur: Real-time ventral metamers for foveated rendering. *ACM Transactions on Graphics* 40, 4 (2021), 1–14.
- [40] Helena Westermann. 2022. Using Chromostereopsis to Enhance Depth Perception in Photos by changing the Hue. (2022).
- [41] Andrew J Woods and Chris R Harris. 2010. Comparing levels of crosstalk with red/cyan, blue/yellow, and green/magenta anaglyph 3D glasses. In *Stereoscopic displays and applications XXI*, Vol. 7524. SPIE, 235–246.
- [42] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proc. of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 2256–2265.
- [43] Haibao Yu, Tuopu Wen, Guangliang Cheng, Jiankai Sun, Qi Han, and Jianping Shi. 2020. Low-bit quantization needs good distribution. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 680–681.
- [44] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing* 26, 7 (2017), 3142–3155.
- [45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of the IEEE conference on Computer Vision and Pattern Recognition*. 586–595.
- [46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of the IEEE International Conference on Computer Vision*. 2223–2232.