Φeat: Physically Grounded Material Feature Representation

Giuseppe Vecchio, Adrien Kaiser, Claudia Cuttano, Romain Rouffet, Rosalie Martin, Elena Garces, and Tamy Boubekeur

Adobe Research, France
{gvecchio, akaiser, ccuttano, rouffet, rmartin, elenag, boubek}@adobe.com

Fig. 1: We present Φeat, a novel physically grounded foundation model sensitive to the physical material properties that govern real-world appearance. We visualize the cosine similarity of Φeat output features between a query patch, marked with a red cross, and all other patches.

Abstract. While foundation models have emerged as general-purpose visual backbones, their representations are primarily optimized for semantics and lack explicit modeling of physical factors, such as reflectance, hindering their efficacy in tasks requiring explicit material reasoning. We introduce **Φeat**, a novel material-grounded visual backbone that encourages a representation sensitive to material identity, including reflectance and mesostructure. Instead of relying on generic data augmentations, we pretrain our model by contrasting observations of the same material under controlled variations in lighting and geometry. This encourages invariance to extrinsic factors while preserving sensitivity to intrinsic material properties. We show that the resulting representation provides strong priors for material-centric tasks, including feature-based material selection and classification. Our results demonstrate that physically inspired weak supervision is an effective strategy for learning representations tailored to material perception.

1 Introduction

Modern computer vision has seen a paradigm shift driven by foundation models, such as the DINO family [11, 36, 45]. These self-supervised models excel at learning representations that are invariant to a wide range of transformations, effectively mapping pixels to a *semantic manifold* where object identity remains stable across different viewpoints and styles. However, while these models capture robust geometric and categorical cues, they often struggle to decouple the physical factors that constitute a scene's appearance, like material properties.

For a representation to truly capture the physical properties that determine the appearance of the world, it must distinguish between **intrinsic factors**, *i.e.*, the physical material properties an object is made of, and **extrinsic factors**, *i.e.*, the context in which that material exists, such as the macro-geometry of the object, the global illumination, and the local orientation. In applications like intrinsic decomposition [23], material capture [15], and robotics, correctly estimating the semantic label (*e.g.*, "chair") is often less critical than the material identity (*e.g.*, "polished oak"). Standard backbones tend to group surfaces by semantic context, whereas material-aware representations should remain stable across changes in shape and lighting for the same intrinsic properties.

The challenge lies in the fact that material identity is not a simple color or texture, but is defined by specific properties, such as reflectance (BRDF), mesostructure, etc. Standard self-supervised objectives, which rely on photometric augmentations like color jittering or solarization, often inadvertently destroy these subtle physical cues or encourage the model to ignore them in favor of object-level consistency. As a result, specialized tasks requiring deep material reasoning still depend on heavy supervision or narrow, domain-specific datasets.

In this paper, we introduce **Φeat**, a visual backbone designed to bridge this gap by specializing a pretrained DINOv3 ViT [45] for **material-aware perception**. Our key insight is that instead of generic data augmentations, we can leverage a physically based supervision. By rendering the same material across a diverse set of geometries and lighting environments, we create "physical triplets" that provide a rigorous training signal for intrinsic invariance. We force the model to associate spatial crops of the same material even when their pixel-level appearance changes dramatically due to extrinsic factors. To do so, we complement the DINO self-supervised training strategy with a cross-material contrastive loss, which further grounds invariance in real physical properties variation rather than semantic similarity. Φeat shifts the focus of the training from object semantics to intrinsic appearance. This produces features that capture underlying material properties while remaining robust to extrinsic perturbations such as local geometry and complex light interactions.

The resulting representation is a material-grounded feature extractor that captures intrinsic reflectance and surface structure while remaining robust to the physical context. Φeat provides a powerful prior for downstream tasks, outperforming general-purpose backbones in material-centric applications without requiring dense per-pixel labels during pretraining.

In a nutshell, our main contributions are:

- **Φeat**, a weakly supervised visual backbone fine-tuned to encode physically grounded material features such as reflectance and geometric mesostructure;
- a material-aware pretraining strategy that leverages synthetic renderings under varied lighting and geometry to encourage invariance to extrinsic appearance factors;
- a large-scale, semantically coherent data generation pipeline, leveraging artist-designed mesh templates to render a vast collection of materials.

2 Related Work

Intrinsic Scene Understanding. The problem of recovering and understanding intrinsic scene characteristics, such as surface reflectance, shape, and illumination, from a single image was formalized as early as the 1970s [3, 30]. This is closely related to material perception, where appearance depends on local optical and physical properties rather than object category alone [1]. Early work on local visual material attributes further showed that material properties can provide discriminative mid-level cues for recognition beyond object and scene context [42]. Despite decades of progress [23], intrinsic scene understanding remains only partially solved. Early progress in intrinsic decomposition was driven by non-learning-based approaches, which produced promising results but struggled to generalize. Two main strategies emerged. The first attempted to classify image gradients as either reflectance or shading, relying on heuristics such as entropy minimization [18] and Retinex-based assumptions [6,28,30], both of which proved brittle in practice. A second line of work extended this idea by clustering regions of similar reflectance [4, 22], typically assuming smooth or continuous lighting within each cluster. Large-scale material recognition was advanced by the Materials in Context database (MINC) [5], which enabled learning-based material classification and segmentation in real-world images. Modern learning-based approaches have achieved remarkable progress in intrinsic decomposition [29, 53] and material capture [15, 33]. However, these methods are typically framed as narrow regression tasks. They rely on large-scale, densely labeled datasets and are often restricted to controlled acquisition setups that reduce ambiguity in the estimation, such as flash illumination [15, 16, 20, 31] or data captured with dedicated scanning devices [21, 40]. Some recent approaches attempt to operate under unknown lighting conditions [33, 50, 51, 57], but they still rely on task-specific supervision. More recently, material segmentation methods [25, 44] have attempted to leverage the representational power of foundation models such as DINO. SAMa [19] similarly targets material-aware selection and segmentation, but in 3D assets, using cross-view consistency to propagate material selections across arbitrary 3D representations. These approaches are highly effective for their target tasks, but typically rely on frozen semantic backbones, additional supervised layers, or task-specific selection pipelines to compensate for the lack

of physical awareness in the underlying features. This reveals a fundamental limitation: existing representations are primarily organized around semantic similarity, providing only a weak foundation for reasoning about materials. In contrast, Φeat learns generic, material-aware features through physically grounded weak supervision. By encoding intrinsic appearance attributes directly into the backbone, our approach enables robust material reasoning that is not tied to rigid physical priors or frozen semantic representations.

Foundation Visual Encoders. Recent advances in visual representation learning are largely driven by large-scale encoders trained either from image–text pairs or from self-supervision. *Vision–language pretraining* learns transferable representations from large collections of image–text pairs. CLIP [38] aligns image and text embeddings through a contrastive objective. SigLIP replaces the softmax formulation with a pairwise sigmoid loss that improves scalability [54], while SigLIP 2 extends this framework with additional training objectives [48]. The Perception Encoder (PE) [8] further shows that contrastive vision–language training yields strong visual representations and exposes different embeddings through alignment strategies: *PE-core* for general-purpose features and *PE Spatial*, which enforces feature consistency within SAM 2 [39] segmentation masks to improve dense prediction. *Self-supervised learning (SSL)* learns representations by enforcing invariance across views of the same image. Methods evolved from pretext tasks [17, 34, 37, 55] to contrastive learning [12, 26] and negative-free variants such as SwAV, BYOL, and SimSiam [10, 13, 24]. More recently, RADIOv2.5 [27] proposed a multi-teacher distillation approach to consolidate these disparate strengths into a single "universal" visual backbone. Building on these ideas, DINO [11] introduced a teacher–student self-distillation framework that learns invariant representations across multiple views of the same image. DINOv2 [36] extends this paradigm by combining global self-distillation with masked patch prediction, producing representations that capture both global semantics and dense spatial structure. DINOv3 [45] further scales this approach and introduces stabilization mechanisms such as Gram anchoring to preserve dense feature structure during large-scale training.

Φeat *physically grounds* these representations by leveraging observations of the same material under varying lighting and geometry. This weak supervision redefines the notion of *similarity* during pretraining, aligning representations with meaningful physical appearance cues.

3 Data Collection and Curation

The effectiveness of foundation model pretraining depends strongly on how the input variability reflects the desired invariances. In Φeat, the goal is to disentangle intrinsic physical properties from extrinsic factors such as geometry and illumination. We design a large synthetic dataset which focuses on variability in lighting and shape while keeping the underlying material identity constant. This controlled variability is what enables the model to learn physically grounded

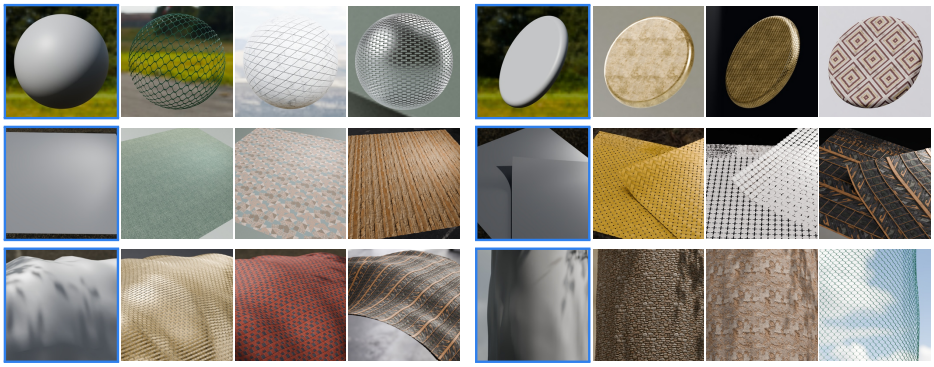

Fig. 2: Examples from our curated synthetic material dataset. Each row shows two geometric templates (highlighted with a blue border) rendered with three different materials each. Templates represent the base geometry onto which materials are applied, while the following images illustrate example renders for different materials on that shape. Templates are semantically aligned with plausible material categories to maintain realistic contexts. The full dataset contains approximately one million high-quality renders generated under diverse lighting, viewpoint, and procedural material variations.

features rather than semantic or contextual similarity, albeit without the need to anchor it in a specific physical model *e.g.*, reflectance distribution function.

Similar datasets have been used for inverse rendering [7, 33, 49] or material similarity tasks by randomly pairing objects, materials, and lighting. This arbitrary pairing introduces two issues: first, it might bias the training toward unrealistic combinations, affecting material appearance [43]; second, it makes large-scale generation impractical. To address these limitations, we carefully build a set of base geometric templates which we pair with semantically meaningful material categories. For example, a cork material is rendered on a rigid, low-curvature surface rather than on a highly wrinkled cloth. This semantic alignment between material and macrogeometry allows the network to focus on contexts where those materials are likely to appear in the real world. Fig. 2 shows different pairing examples of templates and materials.

We rely on the Adobe Substance 3D Assets library [2], which provides over 9,500 procedural materials covering a wide range of 21 appearance classes, including fabric, metal, wood, stone, marble, and plastic. We vary the procedural parameters of these materials following artist-designed presets, to synthesize approximately 36,000 unique sets of PBR (Physically Based Rendering) maps [9]. Each material instance is rendered on a collection of semantically aligned geometries, and illuminated under four high dynamic range environment maps randomly selected from a list of 20, providing diverse lighting conditions. The supplementary materials shows additional materials and the environment maps.

For each render, we randomly vary both the object and environment rotations to ensure a rich sampling of view and light directions. Rendering is per-

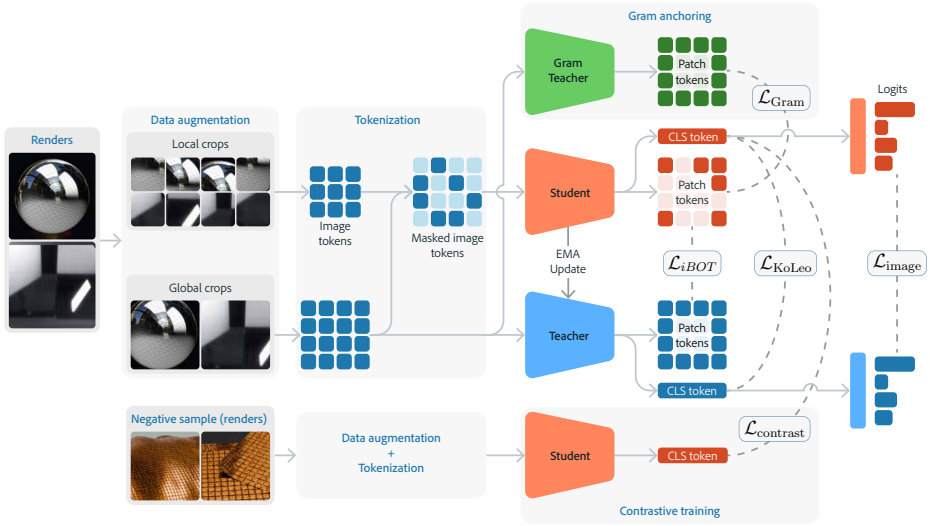

Fig. 3: Φeat training pipeline. Two renderings of the same material are sampled and augmented with a multi-crop strategy that yields global and local views. The *student* processes all crops (globals and locals), with random masking on patch tokens for latent reconstruction; the *teacher* and the *Gram teacher* process *global* crops only. Both networks output class- and patch-level embeddings. At the image level, Sinkhorn-balanced teacher assignments supervise student prototype predictions for *all* student views of the same material, defining $\mathcal{L}_{\text{image}}$. At the patch level, masked student tokens are regressed to the teacher tokens at matching spatial indices, giving $\mathcal{L}_{\text{iBOT}}$. On the student global feature *before* the prototype head, KoLeo encourages dispersion ($\mathcal{L}_{\text{KoLeo}}$) and an in-batch InfoNCE pulls together the two views of the same material while pushing away other materials ($\mathcal{L}_{\text{contrast}}$). Gram anchoring aligns second-order structure on global crops. The teacher is an EMA of the student, and the Gram teacher is a frozen snapshot used only for Gram anchoring.

formed with a Monte Carlo path tracer using physically accurate light transport and microfacet materials [47, 52], ensuring high realism and consistent energy conservation across all scenes. We employ GPU-native path tracing with 128 samples per pixel and apply denoising to all renders. We tessellate each object using the material's displacement map to synthesize realistic self-shadows and inter-reflections. We render the environment map behind transparent regions (holes) to increase variability, helping the model separate thin structures from their surroundings. This process yields roughly one million high-quality renders.

4 Method

Φeat aims to learn visual representations grounded in the intrinsic physical properties of materials. Inspired by recent self-supervised models [36, 45] that rely on image-space augmentations to achieve semantic invariance, we propose a pre-

training strategy centered on **physical invariance**. We replace generic photometric jitter with structured variations of extrinsic factors—rendering the same material across diverse geometries and lighting conditions (Fig. 3). By contrasting these physically grounded views while keeping intrinsic properties constant, Φeat learns to ignore environmental context in favor of material identity.

4.1 Model Architecture

Φeat builds upon the vision transformer (ViT) architecture used in DINOv3 [45]. The network consists of a stack of transformer encoder blocks operating on non-overlapping image patches. Each input image is divided into fixed-size patches of 16×16 pixels, which are flattened and linearly projected into patch tokens. A special global token (often referred to as the `[CLS]` token) is prepended for image-level representation. The model incorporates a small set of learnable "register tokens" that act as dedicated memory slots for global context and mitigate patch-token artifacts. Positional information in the token embeddings is handled via the Rotary Positional Embedding (RoPE) [46] mechanism, which supports token-sequence extrapolation and variable input resolutions. The tokens (global, register, patch) are fed into the transformer encoder stack, which outputs one vector for each token. From the final layer we extract both the patch-level features (one per patch) and the global representation vector from the `[CLS]` token.

4.2 Training Objectives

Φeat follows the self-supervised formulation of DINOv3 [45], extending it with weak supervision on unlabelled physically varying input pairs. Each batch contains 2 global crops and 8 local crops extracted from N renderings $(x_1, ..., x_N)$ of the **same material under different lighting and geometry**. For each crop, the network produces both global and patch-level embeddings. Following DINOv3 [45], we use a combination of image-level objective $\mathcal{L}_{image}$ and a patch-level latent reconstruction objective [56] $\mathcal{L}_{iBOT}$, as well as a $\mathcal{L}_{KoLeo}$ regularizer and Gram anchoring. We complement these objectives by introducing an **in-batch contrastive term** which pulls together representations of the same material while pushing away the embeddings corresponding to different materials. This combination of objectives encourages alignment between physically equivalent renderings while maintaining diversity across unrelated materials.

Image-level objective. The global alignment loss follows the formulation of the DINOv3 objective [45], which replaces the centering and sharpening mechanisms of DINOv2 with the Sinkhorn–Knopp normalization from SwAV [10]. Such normalization of the teacher probabilities enforces balanced assignments over K learnable prototypes, improving stability and preventing feature collapse. Let $f_s(v)$ and $f_t(v)$ be the student and teacher global embeddings for view v, and let $W = [w_1, \dots, w_K]^\top \in \mathbb{R}^{K \times D}$ be the learnable prototype matrix. Student probabilities are

$$p_k(v_s) \;=\; \frac{\exp\!\big(\frac{1}{\tau_s}\, w_k^\top f_s(v_s)\big)}{\sum_{\ell=1}^{K} \exp\!\big(\frac{1}{\tau_s}\, w_\ell^\top f_s(v_s)\big)},$$

while teacher assignments $q_k(v_t)$ are obtained by applying Sinkhorn–Knopp normalization [10] to the teacher logits $\frac{1}{\tau_t} W f_t(v_t)$ over the batch, enforcing balanced use of the K prototypes. Given a set V_t of teacher views and V_s of student views, the cross-entropy objective is

$$\mathcal{L}_{\text{image}} \;=\; -\frac{1}{|V_t|\,|V_s|} \sum_{v_t \in V_t} \sum_{v_s \in V_s} \sum_{k=1}^{K} q_k(v_t)\, \log p_k(v_s). \tag{1}$$

Patch-level latent objective. Following iBOT [56] and DINOv3 [45], a masked latent reconstruction objective is applied at the patch level. A random subset of the student's patch tokens is masked out, and the network is trained to predict the corresponding patch embeddings produced by the teacher on the same spatial positions:

$$\mathcal{L}_{iBOT} = \frac{1}{M} \sum_{m=1}^{M} \left\| h_\theta(p_m^s) - p_m^t \right\|_2^2, \tag{2}$$

where p_m^s and p_m^t denote student and teacher patch embeddings, h_θ is a projection head, and M is the number of masked patches.

KoLeo loss. We use the Kozachenko–Leonenko entropy estimator on the batch of L_2-normalized student features to encourage dispersion and avoid collapse. Let $\{z_i\}_{i=1}^{B}$ be the normalized global student embeddings in a batch and let $\rho_i = \min_{j \neq i} \|z_i - z_j\|_2$ be the nearest-neighbor distance for sample i. Ignoring constants independent of the parameters, maximizing entropy corresponds to minimizing

$$\mathcal{L}_{\text{KoLeo}} \;=\; -\frac{1}{B} \sum_{i=1}^{B} \log\big(\rho_i + \varepsilon\big), \tag{3}$$

with a small $\varepsilon > 0$ for numerical stability. This promotes a more uniform feature distribution on the unit hypersphere.

Gram anchoring. A secondary frozen teacher (Gram teacher) provides a structural target at patch level. We denote by $P_s \in \mathbb{R}^{N \times D}$ and $P_G \in \mathbb{R}^{N \times D}$ the student and Gram-teacher patch matrices after mean-centering across patches. We align second-order patch relations via

$$\mathcal{L}_{\text{Gram}} \;=\; \frac{1}{N^2} \left\| P_s P_s^\top - P_G P_G^\top \right\|_F^2. \tag{4}$$

Contrastive loss. We complement the teacher–student alignment with an in-batch InfoNCE [35] loss on L_2-normalized global student embeddings. For each anchor z_i, positives $P(i)$ are the *other views of the same material*, and all remaining samples in the batch are negatives:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\sum_{j \in P(i)} \exp\big(\text{sim}(z_i, z_j)/\tau\big)}{\sum_{k \neq i} \exp\big(\text{sim}(z_i, z_k)/\tau\big)}, \tag{5}$$

where sim is cosine similarity and τ is a temperature. This objective explicitly pulls together representations of the same physical material, while pushing apart embeddings corresponding to different materials. It complements the DINO objective by reinforcing instance-level consistency independently of the teacher distribution, anchoring the model's global features to material identity.

Total objective. The total loss consists of a weighted sum of these components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{image}} + \lambda_{\text{p}}\,\mathcal{L}_{iBOT} + \lambda_{\text{k}}\,\mathcal{L}_{\text{KoLeo}} + \lambda_{\text{g}}\,\mathcal{L}_{\text{Gram}} + \lambda_{\text{c}}\,\mathcal{L}_{\text{contrast}}, \tag{6}$$

with the coefficients reported in Section 5.

The combination of our augmentation strategy and the cross-material contrastive loss shifts the inductive bias of the pretraining from semantic consistency across image crops to physical invariance across varying extrinsic conditions.

5 Experiments

We evaluate whether Φeat learns representations that are invariant to extrinsic factors, such as lighting and geometry, while remaining sensitive to intrinsic material properties. We consider three main evaluations: **material selection**, based on patch-level material similarity; **feature separability**, measured with a non-parametric k-nearest neighbors (k-NN) classifier; and **robustness**, measured through prediction stability under illumination and geometry changes. Additionally, we assess the **semantic–material tradeoff** by evaluating how much general-purpose semantic performance is retained after material-centric adaptation. Together, these experiments test whether the learned features capture material identity independently of environmental conditions and quantify the compromise on semantic capabilities.

Baselines. We compare Φeat against strong vision foundation models trained with different supervision paradigms. We consider CLIP [38] and SigLIP 2 [48] as vision–language models trained on large-scale image–text pairs, the Perception Encoder [8] in both its PE-core and PE Spatial variants, the agglomerative RADIOv3 [27], and the self-supervised models DINOv2 with registers [14, 36] and DINOv3 [45]. These models represent the strongest publicly available visual encoders. All methods use their *ViT-B* backbones, and we adopt input resolutions corresponding to 1024 patch tokens (448×448 for patch size 14, 512×512 for patch size 16) for a fair comparison.

Implementation Details. We train Φeat for 10,000 iterations using batches of 512 render pairs, each consisting of two different views of the same material. Each pair is augmented using a multi-crop strategy, producing 2 global and 8 local crops per view. The global and local views are extracted by randomly cropping 40–100% and 10–40% of the renders respectively, ensuring a diverse distribution of context and texture. We use a *ViT-B/16* backbone initialized from DINOv3 weights and optimized using AdamW [32] with a base learning rate of 0.001 and weight decay of 0.05. The teacher momentum follows a cosine schedule between 0.996 and 1.0, and the student temperature τ_s is fixed to 0.1.

Table 1: Material selection and classification results. We compare Φeat with state-of-the-art foundation visual encoders trained with different supervision signals. Pixel-level metrics evaluate material selection performance, while classification metrics assess feature separability using k-NN retrieval. Φeat achieves the strongest overall performance, improving material selection and classification across metrics, indicating that the learned representation is both physically grounded and discriminative. Best results **bold**, 2nd best underlined. Task-specific supervised material selection methods are reported as supervised references.

Model	Material Selection			k-NN Classification		
	$\ell_1 \downarrow$	mIoU ↑	F1 ↑	Acc. ↑	Prec. ↑	F1 ↑
Supervised (upper bound)						
Materialistic [44]	5.7	85.8	90.6	—	—	—
Guerrero et al. 2025 [25]	3.0	89.6	93.5	—	—	—
Weakly supervised						
CLIP [38]	55.7	21.6	31.6	39.9	36.6	31.2
SigLIP 2 [48]	53.1	24.4	35.2	33.5	30.8	25.1
PE-core [8]	37.3	40.5	54.6	44.2	37.0	33.9
Agglomerative						
PE Spatial [8]	**25.2**	61.4	72.3	50.3	44.0	39.0
RADIOv3 [27]	25.6	<u>63.5</u>	<u>74.7</u>	35.9	32.9	28.8
Self-supervised						
DINOv2 [36]	26.7	56.5	69.3	64.3	56.9	53.3
DINOv3 [45]	27.6	59.7	72.2	<u>66.9</u>	<u>60.0</u>	<u>55.4</u>
Φeat (ours)	<u>25.4</u>	**72.0**	**80.8**	**71.8**	**75.0**	**66.6**

We train on input resolutions of 224^2 for global crops and 112^2 for local crops, following the multi-crop strategy described in Section 4. Training is performed on 16 NVIDIA A100 GPUs using mixed-precision distributed data parallelism, for a total approximate time of 100 hours. We set the loss weights to $\lambda_{\text{p}} = 1.0$, $\lambda_{\text{k}} = 0.1$, $\lambda_{\text{g}} = 0.7$, and $\lambda_{\text{c}} = 0.25$. For the KoLeo term we use $\varepsilon = 10^{-6}$ in Eq. (3); for the patch objective we randomly mask between 10% and 50% of the student patches with a probability of 50%; the InfoNCE temperature is fixed to $\tau = 0.1$. We enable the Gram anchoring for the last 2,000 iteration steps.

5.1 Quantitative Results

We evaluate the representations learned by Φeat on two downstream tasks: **material selection** and **feature separability** via k-NN. Together, these assess whether the learned features capture intrinsic appearance properties independently of geometry and lighting. Additional qualitative results, including material selection on DuMaS and patch similarity in images and videos through cross-frame feature propagation, are provided in the supplementary materials.

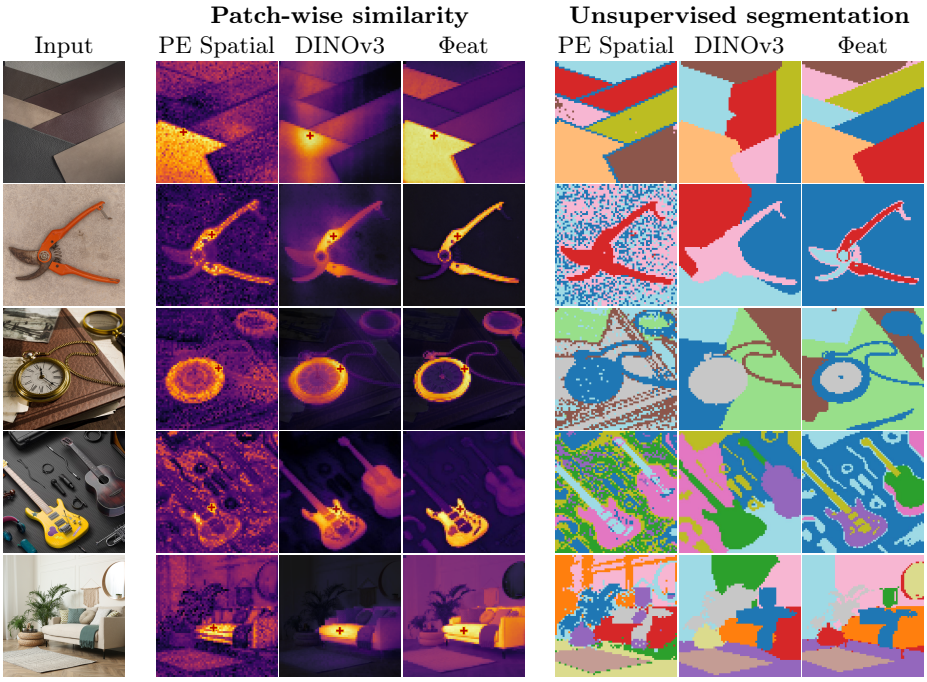

Fig. 4: Patch-wise similarity and unsupervised segmentation. Left group shows cosine similarity maps between the embedding of a reference patch (red cross) and all others, visualizing the spatial coherence of learned representations. The examples shown gradually evolve from a flat surface to a medium-scale scene. Right group displays K-means segmentations obtained from the patch embeddings. Compared to DINOv3 and PE Spatial, Φeat produces similarity responses and clusters that are more spatially consistent toward material properties, grouping regions by reflectance and texture rather than by semantic or geometric cues, making it suitable for appearance-related tasks.

Material selection. We follow the evaluation protocol and DuMaS dataset of [25] to assess whether features separate materials according to appearance similarity. Given a query patch, we compute its cosine similarity to all image-patch embeddings, producing a dense similarity map. This map is thresholded at 0.5 and compared against the ground-truth material segmentation. As shown in Table 1, Φeat consistently outperforms all generic feature baselines, including DINOv3 [45], from which it is initialized. Agglomerative models such as PE Spatial [8] and RADIOv3 [27] provide competitive dense-prediction features, but remain less aligned with physical material boundaries. Conversely, vision-language models such as CLIP [38] and SigLIP 2 [48] perform poorly, as their embeddings tend to conflate material identity with object-level semantics. Overall, Φeat produces more coherent similarity maps that better follow material boundaries while remaining robust to changes in illumination and geometry. We also report performance for Materialistic [44] and Guerrero et al. [25] as task-

specific upper bounds. Unlike these methods, which use pixel-level supervision and dedicated dense-prediction architectures, Φeat is evaluated directly from ViT patch-level backbone features, without supervised task adaptation. Thus, in downstream pipelines, Φeat is intended to replace generic backbones such as DINOv3 rather than full task-specific prediction architectures.

Feature separability via k-NN evaluation. To assess whether the learned representations organize materials into semantically meaningful clusters, we adopt a non-parametric k-nearest neighbors (k-NN) evaluation. We use a controlled synthetic test set containing 972 materials within 16 categories. Each material is rendered under 6 geometry templates and 4 lighting conditions, producing 24 distinct variants per material, resulting in a total of $23,328$ renders.

Each image is embedded with the frozen encoder. For models without a global token, such as *PE Spatial* [8], we use global average pooling over spatial features. For each query sample, we retrieve its $k = 16$ nearest neighbors and assign a label through weighted majority voting, where weights are proportional to the cosine similarity. Classification accuracy is then computed with respect to the ground-truth material labels. As shown in Table 1, Φeat significantly outperforms all baselines, including DINOv3. These results show that the physically oriented pretraining of Φeat produces tighter, more homogeneous clusters that are grounded in intrinsic material properties. Interestingly, models optimized for dense prediction or universal distillation, such as PE Spatial and RADIOv3, struggle in this global classification setting, further highlighting that Φeat captures a more robust physical material identity than semantic or multi-purpose backbones.

Robustness. To assess invariance to extrinsic factors, we measure robustness across variations in illumination and geometry. For each material class we compute predictions using the same k-NN classifier described above. We then evaluate the average pairwise Hamming distance between predictions obtained under different lighting conditions (fixed geometry) and under different geometry templates (fixed lighting). Lower values indicate higher invariance. Table 3 shows that Φeat outperforms the other methods, exhibiting less variability under both types of perturbation and confirming that the representation remains stable across extrinsic appearance changes.

5.2 Semantic Performance

We quantify the tradeoff between material alignment and general-purpose utility in Tab. 2 by evaluating Φeat on the original semantic tasks via linear probing: ImageNet classification and ADE20K/Cityscapes segmentation. Compared to DINOv3, Φeat incurs a moderate drop of 3.4 accuracy points for classification and 5.3/5.4 mIoU points for segmentation, respectively. This indicates that the material-centric adaptation **preserves much of DINOv3's semantic** utility while reorienting the representation toward material identity. We therefore view this as a **specialization tradeoff**: DINOv3 remains stronger for purely se-

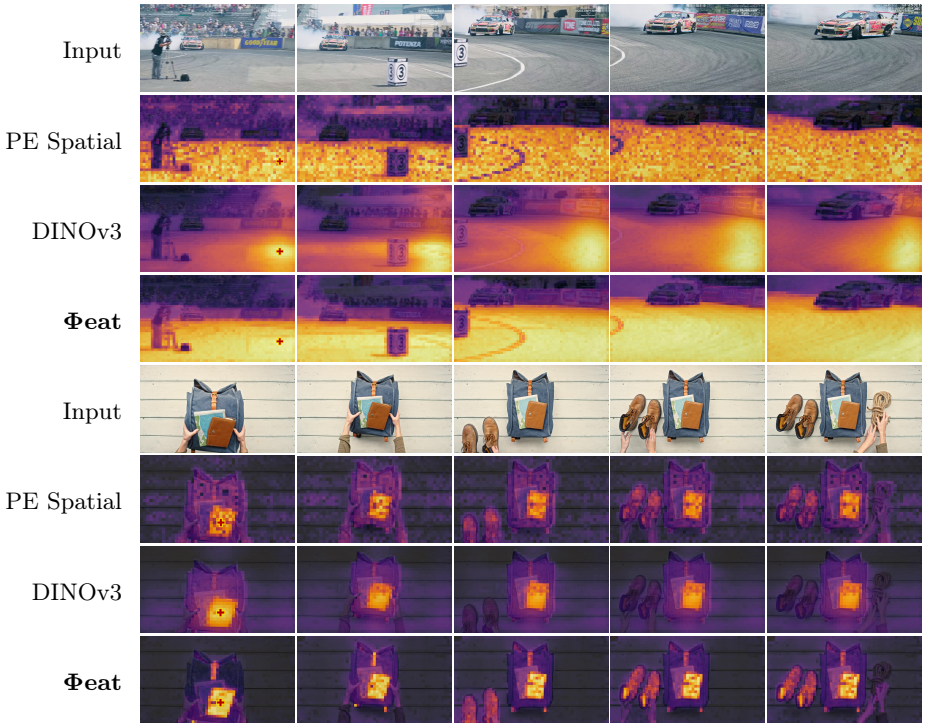

Fig. 5: Feature similarity in videos. The selection is initialized in the first frame and propagated throughout the video. We compute the cosine similarity maps between the embedding of a reference patch (red cross) and all patches of all frames in the video. Over time, Φeat maintains better coherence and alignment with material boundaries compared to DINOv3 and PE Spatial. In the first video, Φeat more uniformly selects the asphalt compare to the other methods, while in the second it identifies all the leather parts regardless of the object semantics.

mantic tasks, while Φeat provides more material-grounded features for material-consistency and appearance-based tasks.

5.3 Qualitative Results

We qualitatively compare Φeat against DINOv3 [45] and PE Spatial [8] using patch-wise similarity and unsupervised segmentation (Fig. 4). Each heatmap shows cosine similarity between a reference patch, marked with a cross, and all other image patches. DINOv3 often highlights semantically or spatially related regions, typically parts of the same object, even when materials differ, while PE Spatial produces noisier responses. In contrast, Φeat yields coherent similarity maps aligned with material boundaries, grouping patches by reflectance and texture rather than object identity. Fig. 5 shows cross-frame selection in videos, where the selection is initialized in the first frame and propagated over time.

Table 2: Semantic–material tradeoff vs. DINOv3. We evaluate the semantic performance retained by Φeat after material-centric adaptation using linear probing on ImageNet classification (acc, ↑) and ADE20K/Cityscapes semantic segmentation (mIoU, ↑). Although Φeat incurs a moderate drop compared to the original DINOv3 backbone, it preserves most of its semantic utility while reorienting the representation toward material-aware features.

	Classification	**Semantic Segmentation**	
	ImageNet	ADE20K	Cityscapes
	Acc. ↑	mIoU ↑	
DINOv3	83.5	51.5	71.5
Φeat (**ours**)	80.1	46.2	66.1

Table 3: Robustness to illumination and geometry variations, measured as the avg. pairwise Hamming distance between k-NN predictions across renderings of the same material under different lighting or geometry. Lower values indicate higher invariance.

	Illumination	**Geometry**
	hamming ↓	hamming ↓
CLIP [38]	0.403	0.534
DINOv2 [36]	0.284	0.435
DINOv3 [45]	0.240	0.365
Φeat (ours)	**0.221**	**0.305**

Table 4: Ablation study. Effect of each training component of Φeat starting from the DINOv3 backbone. Multi-render supervision and the contrastive objective progressively improve both material selection and k-NN classification performance, leading to stronger material grouping.

	Material selection			**k-NN**
	ℓ_1 ↓	IoU ↑	F_1 ↑	Acc. ↑
DINOv3	27.6	59.7	72.2	66.9
+ single render	26.5	69.4	78.1	34.5
+ multi render	26.1	69.9	79.2	51.3
+ contrastive	**25.4**	**72.0**	**80.8**	**71.8**

Compared to DINOv3 and PE Spatial, Φeat better preserves coherence and alignment with material boundaries across frames.

For unsupervised segmentation, we apply K-means clustering to patch embeddings to visualize how features organize the image space. The number of clusters is selected automatically by evaluating $K \in \{2, \ldots, 12\}$ and choosing the value that maximizes the silhouette score [41]. Segmentations from DINOv3 and PE Spatial often follow semantic cues, grouping object parts while mixing distinct materials. Φeat instead produces spatially consistent and physically meaningful clusters that better separate regions according to material properties.

These qualitative results show that Φeat learns representations driven by reflectance and texture rather than semantics, capturing physically grounded appearance features robust to geometry and lighting. Additional results showing invariance to lighting are provided in the supplementary materials.

5.4 Ablation Study

We conduct an ablation study to evaluate the effect of each novel training component. Table 4 shows that fine-tuning with single-render material supervision already improves the material-selection metrics (ℓ_1, IoU, F_1), but it also substantially reduces k-NN accuracy, showing that the model becomes more physically grounded while losing global discriminability. Adding the *multi-render* scheme improves both material-selection and classification performance, suggesting that exposure to varied viewpoints and illumination stabilizes the representation. However, only the contrastive term fully resolves the loss of separability introduced by material supervision, enforcing compact intra-material clusters and clearer inter-material boundaries. This combination yields the strongest performance across all metrics and consistently surpasses DINOv3, leading to a material representation that is both physically grounded and discriminative.

6 Limitations

While Φeat successfully learns invariance to extrinsic factors, it does not explicitly disentangle the latent space into interpretable physical factors. A more expressive formulation separating roughness, metalness, or albedo would allow users to manipulate features along specific physical dimensions. Moreover, our pretraining relies entirely on synthetic data; closing the domain gap by integrating unlabelled real-world data through domain adaptation or self-training remains an open challenge. As a result of its material-centric adaptation, Φeat also incurs a moderate loss on general-purpose semantic tasks compared to the original backbone, reflecting a specialization tradeoff between semantic utility and material alignment. Finally, our current physical model is primarily optimized for surface reflectance and shading. Modeling more complex light-transport phenomena, such as translucency and subsurface scattering, could further improve its physical fidelity.

7 Conclusion

We proposed Φeat, a weakly supervised visual backbone that learns physically grounded representations of appearance directly from unlabelled data. By replacing photometric augmentations with physically meaningful variations *i.e.*, different renderings of the same material under diverse geometries and lighting conditions, our approach bridges the gap between semantic invariance and physical material understanding. Through a combination of DINO-based teacher–student alignment and contrastive regularization, Φeat captures material identity while remaining invariant to extrinsic factors such as shape and illumination. We demonstrated that the learned features transfer effectively to downstream material tasks, including fine-grained material selection, unsupervised segmentation and clustering under changing illumination and geometry. These results confirm that material-aware pretraining can yield representations that are both robust and interpretable, without relying on semantic labels.

References

1. Adelson, E.H.: On seeing stuff: the perception of materials by humans and machines. In: Human vision and electronic imaging VI. vol. 4299, pp. 1–12. SPIE (2001)
2. Adobe: Substance 3D Assets. `https://substance3d.adobe.com/assets` (2025)
3. Barrow, H.G., Tenenbaum, J.M.: Recovering intrinsic scene characteristics. In: Hanson, A.R., Riseman, E.M. (eds.) Computer Vision Systems, pp. 3–26. Academic Press (1978)
4. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM Transactions on Graphics (TOG) **33**(4), 1–12 (2014)
5. Bell, S., Upchurch, P., Snavely, N., Bala, K.: Material recognition in the wild with the materials in context database. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3479–3487 (2015)
6. Bi, S., Han, X., Yu, Y.: An l_1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. ACM Transactions on Graphics (TOG) **34**(4), 1–12 (2015)
7. Birsak, M., Femiani, J., Zhang, B., Wonka, P.: MatCLIP: Light- and shape-insensitive assignment of PBR material models. In: ACM SIGGRAPH 2025 Conference Papers. pp. 1–10 (2025)
8. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. arXiv preprint arXiv:2504.13181 (2025)
9. Burley, B.: Extending the Disney BRDF to a BSDF with integrated subsurface scattering. SIGGRAPH 2015 Course: Physically Based Shading in Theory and Practice (2015), `https://api.semanticscholar.org/CorpusID:208625014`
10. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
11. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
12. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
13. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
14. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
15. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Single-image SVBRDF capture with a rendering-aware deep network. ACM Transactions on Graphics (TOG) **37**(4), 1–15 (2018)
16. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Flexible SVBRDF capture with a multi-image deep network. In: Computer Graphics Forum. vol. 38, pp. 1–13. Wiley Online Library (2019)
17. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE international conference on computer vision. pp. 1422–1430 (2015)

18. Finlayson, G.D., Drew, M.S., Lu, C.: Intrinsic images by entropy minimization. In: European conference on computer vision. pp. 582–595. Springer (2004)
19. Fischer, M., Georgiev, I., Groueix, T., Kim, V.G., Ritschel, T., Deschaintre, V.: SAMa: Material-aware 3D selection and segmentation. In: 2026 International Conference on 3D Vision (3DV). pp. 1812–1822. IEEE (2026)
20. Gao, D., Li, X., Dong, Y., Peers, P., Xu, K., Tong, X.: Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images. ACM Transactions on Graphics (TOG) **38**(4), 1–17 (2019)
21. Garces, E., Arellano, V., Rodriguez-Pardo, C., Pascual-Hernandez, D., Suja, S., Lopez-Moreno, J.: Towards material digitization with a dual-scale optical system. ACM Transactions on Graphics (TOG) **42**(4), 1–13 (2023)
22. Garces, E., Munoz, A., Lopez-Moreno, J., Gutierrez, D.: Intrinsic images by clustering. In: Computer graphics forum. vol. 31, pp. 1415–1424. Wiley Online Library (2012)
23. Garces, E., Rodriguez-Pardo, C., Casas, D., Lopez-Moreno, J.: A Survey on Intrinsic Images: Delving Deep into Lambert and Beyond. International Journal of Computer Vision **130**(3), 836–868 (2022)
24. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
25. Guerrero-Viu, J., Fischer, M., Georgiev, I., Garces, E., Gutierrez, D., Masia, B., Deschaintre, V.: Fine-grained spatially varying material selection in images. ACM Transactions on Graphics (TOG) **44**(6), 1–11 (2025)
26. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
27. Heinrich, G., Ranzinger, M., Yin, H., Lu, Y., Kautz, J., Tao, A., Catanzaro, B., Molchanov, P.: RADIOv2.5: Improved baselines for agglomerative vision foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22487–22497 (2025)
28. Horn, B.K.: Determining lightness from an image. Computer graphics and image processing **3**(4), 277–299 (1974)
29. Kocsis, P., Sitzmann, V., Nießner, M.: Intrinsic image diffusion for indoor single-view material estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5198–5208 (2024)
30. Land, E.H., McCann, J.J.: Lightness and retinex theory. Journal of the Optical society of America **61**(1), 1–11 (1971)
31. Li, Z., Sunkavalli, K., Chandraker, M.: Materials for masses: SVBRDF acquisition with a single mobile phone image. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 72–87 (2018)
32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
33. Martin, R., Roullier, A., Rouffet, R., Kaiser, A., Boubekeur, T.: MaterIA: Single image high-resolution material capture in the wild. In: Computer Graphics Forum. vol. 41, pp. 163–177. Wiley Online Library (2022)
34. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016)
35. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)

36. 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)
37. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
38. 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)
39. 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 (2024)
40. Rodriguez-Pardo, C., Casas, D., Garces, E., Lopez-Moreno, J.: TexTile: A differentiable metric for texture tileability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4439–4449 (2024)
41. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
42. Schwartz, G., Nishino, K.: Automatically discovering local visual material attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3565–3573 (2015)
43. Serrano, A., Chen, B., Wang, C., Piovarci, M., Seidel, H.P., Didyk, P., Myszkowski, K.: The effect of shape and illumination on material perception: model and applications. ACM Transactions on Graphics (TOG) **40**(4), 1–16 (2021). `https://doi.org/10.1145/3450626.3459813`
44. Sharma, P., Philip, J., Gharbi, M., Freeman, B., Durand, F., Deschaintre, V.: Materialistic: Selecting similar materials in images. ACM Transactions on Graphics (TOG) **42**(4), 1–14 (2023). `https://doi.org/10.1145/3592390`
45. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
46. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: RoFormer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
47. Trowbridge, S., Reitz, K.P.: Average irregularity representation of a rough ray reflection. Journal of the Optical Society of America **65**(5), 531–536 (1975)
48. 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)
49. Vecchio, G., Deschaintre, V.: MatSynth: A modern PBR materials dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22109–22118 (June 2024)
50. Vecchio, G., Martin, R., Roullier, A., Kaiser, A., Rouffet, R., Deschaintre, V., Boubekeur, T.: ControlMat: A controlled generative approach to material capture. ACM Transactions on Graphics **43**(5), 1–17 (2024)
51. Vecchio, G., Palazzo, S., Spampinato, C.: SurfaceNet: Adversarial SVBRDF estimation from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12840–12848 (2021)

52. Walter, B., Marschner, S.R., Li, H., Torrance, K.E.: Microfacet models for refraction through rough surfaces. In: Proceedings of the 18th Eurographics conference on Rendering Techniques. pp. 195–206 (2007)
53. Zeng, Z., Deschaintre, V., Georgiev, I., Hold-Geoffroy, Y., Hu, Y., Luan, F., Yan, L.Q., Hašan, M.: RGB↔X: Image decomposition and synthesis using material- and lighting-aware diffusion models. In: ACM SIGGRAPH 2024 Conference Papers (2024). https://doi.org/10.1145/3641519.3657445
54. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
55. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision. pp. 649–666. Springer (2016)
56. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: iBOT: Image BERT pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)
57. Zhou, X., Kalantari, N.K.: Adversarial single-image SVBRDF estimation with hybrid training. In: Computer Graphics Forum. vol. 40, pp. 315–325. Wiley Online Library (2021)