Attentive Illumination Decomposition Model for Multi-Illuminant White Balancing

Dongyoung Kim[1] Jinwoo Kim[1] Junsang Yu[2] Seon Joo Kim[1]
[1]Yonsei University [2]Samsung Advanced Institute of Technology

Abstract

White balance (WB) algorithms in many commercial cameras assume single and uniform illumination, leading to undesirable results when multiple lighting sources with different chromaticities exist in the scene. Prior research on multi-illuminant WB typically predicts illumination at the pixel level without fully grasping the scene's actual lighting conditions, including the number and color of light sources. This often results in unnatural outcomes lacking in overall consistency. To handle this problem, we present a deep white balancing model that leverages the slot attention, where each slot is in charge of representing individual illuminants. This design enables the model to generate chromaticities and weight maps for individual illuminants, which are then fused to compose the final illumination map. Furthermore, we propose the centroid-matching loss, which regulates the activation of each slot based on the color range, thereby enhancing the model to separate illumination more effectively. Our method achieves the state-of-the-art performance on both single- and multi-illuminant WB benchmarks, and also offers additional information such as the number of illuminants in the scene and their chromaticity. This capability allows for illumination editing, an application not feasible with prior methods.

1. Introduction

Color constancy, a unique human capability, allows us to perceive the color of objects uniformly under any lighting conditions. Similarly, a computational color constancy or white balancing (WB) module is integrated into the image processing unit, designed to compensate for the effects of illumination to recover the original color of the objects.

Many WB studies have been conducted with the goal of predicting the single chromaticity vector of the light source for a given image, assuming uniform illumination. Traditional statistics-based methodologies [16, 19, 20, 43], including gray world [10] and white patch [30] algorithms, used various statistics that could be obtained from images. Data-driven methods [4, 24] worked by optimizing through the white balance dataset. However, these algorithms produce distorted results when multiple illuminants affect the scene simultaneously. For example, when a blue skylight is coming in from the window into a room with warm-colored lighting, applying a single white balance matrix to the entire image may fail in recovering the scene color.

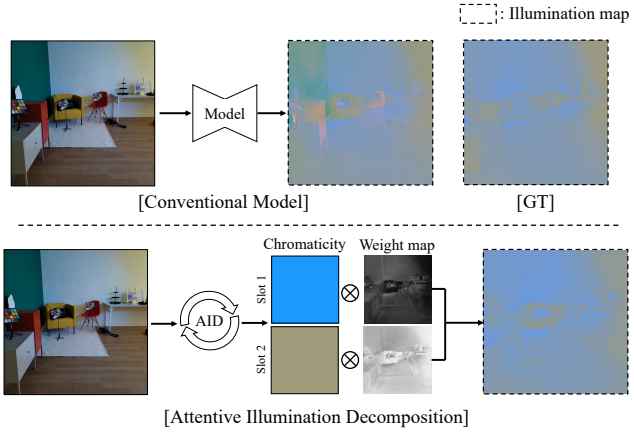

Figure 1. Comparison of the AID framework (bottom) with existing approaches (top). Previous methodologies did not individually consider illuminant profiles within the scene, resulting in unnatural results. The AID framework outperforms previous works in illumination estimation by estimating the chromaticity and pixel-wise weight map of each individual illuminant and combining them.

Accordingly, spatially varying white balance algorithms have been proposed to deal with multi-illuminant scenes. Early works estimate mixed illumination map by utilizing auxiliary flash photography [25] or prior knowledge about the chromaticity of illuminants [5, 23]. Recently, many DNN-based methods have been introduced with the advancements in neural networks. Algorithms using patches [8], GANs [42], U-Net [27] with transformer blocks [31] have been proposed.

All previous multi-illuminant WB works directly generate patch- or pixel-level predictions of illumination maps

Author E-mail : dongyoung.kim@yonsei.ac.kr

using an encoder-decoder structure without any structural constraints. These approaches often fail to satisfy the linearity constraint [21, 23, 27] that the chromaticity of mixed illumination can be expressed as a linear combination of individual light source chromaticity under the Lambertian image model. This may result in producing unnatural illumination that does not exist in a scene (Fig. 1 top). In addition, as the previous methods cannot offer individual light source profiles in a multi-illuminant scene, further tuning or editing the illumination is not possible.

To overcome the limitations of the existing multi-illuminant WB methods, we propose the Attentive Illumination Decomposition (AID) mechanism. AID shows strong performance and is equipped with tunability for the spatially varying multi-illuminant WB. Our framework works in an end-to-end manner with a single given image. In other words, it does not require any auxiliary images [25] or post-processing procedures [25, 26] to decompose the illumination map. Our model is based on slot attention [33], to learn the implicit representation of illuminant chromaticity in a scene in the form of slot vectors. Specifically, we leverage the slot vectors to represent the chromaticities of the light sources in a scene, and use the attention map of each slot as the pixel-wise weight map of corresponding illuminant (Fig. 1 bottom). By doing so, we can enforce each predicted pixel-wise chromaticity to be a linear combination of the slot chromaticities, and enable illuminant-wise tunability. The way our model generates the final illumination maps follows the linearity constraint so that our method can properly tackle the problem of spatially varying WB. Furthermore, we propose a novel loss called centroid-matching loss, to effectively train our slot attention based model by assigning specific color ranges to slots.

We validate the robustness of AID framework through comprehensive experiments conducted on various datasets, including the LSMI dataset [27], the Multi-Illumination In the Wild dataset [34], and the well-established single-illuminant dataset, NUS-8 [11]. The experimental results consistently demonstrate superior performance compared to previous models, achieving the state-of-the-art performance across all of the aforementioned datasets.

Our contributions can be summarized as follows:

- By successfully leveraging the concept of the slot attention, we propose a novel end-to-end framework AID, which can infer the chromaticities of illuminants and their pixel-level weight maps separately.
- We introduce the centroid-matching loss to enable more effective updates of slots to represent specific color gamuts.
- Our model not only demonstrates the state-of-the-art performance in both single- and multi-illuminant white balance scenarios but also provides tunable WB, thanks to its capacity to generate fully disentangled illumination maps.

2. Related work

2.1. Computational color constancy

Single-illuminant WB. Classical statistics-based algorithms utilizing image statistics have been studied [10, 15, 30, 43] for computational color constancy. Additionally, numerous WB datasets [11, 13, 17, 40] have been proposed for data-driven research. Methodologies have been introduced involving the learning of kernels to detect illuminant chromaticity in the uv-histogram space [3, 4], utilizing convolutional features [7, 24, 36, 41], and employing various learning techniques [32, 37, 45, 46]. In particular, FC4 [24] employs a form of attention technique by inferring spatial weighting coefficients, rather than uniformly using all spatial features within the image. On the other hand, C4 [46] demonstrated the capability for more accurate chromaticity inference through iterative refinement process. While they achieve impressive results for single-illuminant WB, they cannot address the multi-illuminant WB cases. We found that the incorporation of spatial attention maps and an iterative refinement strategy, in conjunction with the concept of slots outlined in Sec. 2.2, is highly suitable for addressing the spatially varying multi-illuminant decomposition task.

Multi-illuminant WB. To solve the multi-illuminant WB problem, several studies have been conducted to utilize additional prior information such as the chromaticity of the illuminant [5, 23], flash photography [25], and human face [6]. Approaches that apply single-illuminant WB in a spatially varying form have been introduced in [9, 21], and a graph structure reflecting the characteristics of spatially varying WB has been utilized in [35].

Following small-scale multi-illuminant datasets [5, 8, 9, 21] for testing spatially varying WB algorithms, several large scale multi-illuminant datasets have been captured [27, 34] and synthesized [22] recently. Deep learning-based strategies such as using GANs [42], and leveraging transformer blocks with multi-task learning strategies [31] have also been explored.

The base architecture for previous multi-illuminant WB used the encoder-decoder structure to directly predict the chromaticity of illumination for each individual pixel. These models fall short in estimating and incorporating the individual chromaticities of illuminants present in the scene, leading to inconsistencies in the generated illumination map. (Fig. 1 top). While a model that estimates pixel-wise weights for pre-specified WB presets [2] has been proposed recently, the resulting weight maps do not accurately reflect the the ground-truth illuminant-wise mixing ratio due to its reliance on pre-defined WB presets. A summary of the comparison between our framework and previous works is presented in Table 1.

Models	Mixed illumination	Decomposed illumination map	Controllable WB
Single AWB	×	×	×
Multi-AWB [2, 31, 42]	✓	×	×
AID (Ours)	✓	✓	✓

Table 1. Comparison between previous WB methods and our AID framework. AID predicts a decomposed illumination map, enabling the inference of individual illuminant chromaticity and the number of illuminants in a scene. This new feature enables controllable WB, allowing for individual adjustment of the color of each illuminant in a scene.

2.2. Slot attention

Slot attention [33] was introduced to solve the object-centric learning (OCL), where the model needs to cluster and compute the representation of objects from a given scene without any human-annotated labels in an autoencoding manner. Slot attention employs the concept of the slots, a set of vectors, each of which captures the representation of the object in a scene. Slots are initialized using Gaussian random sampling and are subsequently evolved to capture the representations of objects. Dot-product based attention maps between slots and encoded visual feature maps are used for updating the slots. By applying slot-wise softmax mechanism on the attention map, slots are forced to compete with each other to get more task-relevant representation, i.e. object-centric representation.

Due to the decomposition ability of the slot attention, it has been widely applied to various domains in computer vision such as object discovery [14, 28, 29, 33], novel view synthesis [39], panoptic segmentation [47], and visual question answering [44]. Slot attention acts like a soft k-means clustering, where the slots are appropriately updated to represent the target sub-element. In this work, we adopt slot attention to the task of multi-illuminant white balancing, enforcing slots to implicitly represent individual illuminants. In addition, we introduce a novel loss function named centroid matching loss, aimed at preventing all slots from indiscriminately contributing to the inference. This improves illumination decomposition accuracy by allocating the specific color ranges to each slot.

3. Method

3.1. Image formation model

In the Lambertian image model, the RGB value of each pixel under single-illuminant condition can be represented as follows:

$$I = k\boldsymbol{\rho} \circ \boldsymbol{\ell}. \tag{1}$$

Here I, $\boldsymbol{\rho}$ and $\boldsymbol{\ell}$ are 3×1 vectors for observed RGB pixel, surface reflectance, and normalized illuminant chromaticity, respectively. The $\circ$ symbol represents element-wise product and the scalar k represents the integrated scaling term of illumination including the power of illuminant and surface normal. In this paper, we normalize the illuminant chromaticity so that the value of the green channel becomes 1. Previous works [21, 23, 27] suggest that if multiple illuminants are present in a scene, the chromaticity of mixed illumination can be represented by the linear combination of the chromaticity of each illuminant. This property has been used to calculate per-pixel illumination labels for multi-illuminant datasets [5, 27].

Under the imaging model, the illumination chromaticity value $\boldsymbol{\ell}$ on a given location x of a single or multiple illuminant scene can be generalized and expressed as follows:

$$\begin{aligned} \boldsymbol{\ell}_{mixed}(x) &= \sum_{k=1}^{N} \alpha_k(x)\boldsymbol{\ell}_k, \\ \text{where} \quad \sum_{k=1}^{N} \alpha_k(x) &= 1, \quad \boldsymbol{\ell}_k = \begin{bmatrix} R_k \\ B_k \end{bmatrix}. \end{aligned} \tag{2}$$

α_k and $\boldsymbol{\ell}_k$ represent the weight map and the normalized chromaticity of illuminant k, respectively, and N is the number of illuminants in the scene. As mentioned earlier, we only consider the R and B channels of the illuminant chromaticity $\boldsymbol{\ell}_k$, given that the G channel is normalized to 1. Since $\boldsymbol{\ell}_k$ is the chromaticity of each light source, it does not change with pixel location, and only the weight map α is dependent on x.

3.2. Attentive illumination decomposition

To solve the multi-illuminant WB, we design the Attentive Illumination Decomposition (AID) framework, which follows the imaging model described in Eq. (2). The proposed method first predicts the weight map α_k and the chromaticity $\boldsymbol{\ell}_k$ of each illuminant in the scene, and then the mixed illumination map $\boldsymbol{\ell}_{mixed}$ for WB is generated using Eq. (2). It is important to highlight that this is the first approach to separately predict the chromaticity and the weight map of illuminant for multi-illuminant WB, leading to enhancement in performance.

To obtain α_k and $\boldsymbol{\ell}_k$, our framework utilizes the slot attention [33]. Different from the existing slot attention models, where slots are typically designed to capture object-level features, we design the model to make the slots to represent illuminant-level information. More precisely, each slot in our model allows us to infer both the chromaticity and the weight map associated with the corresponding illuminant. The overview of our framework is illustrated in Fig. 2(a). Our model consists of three parts: 1) image feature extraction, 2) iterative slot calibration process using slot attention, and 3) weight map & illuminant chromaticity fusion.

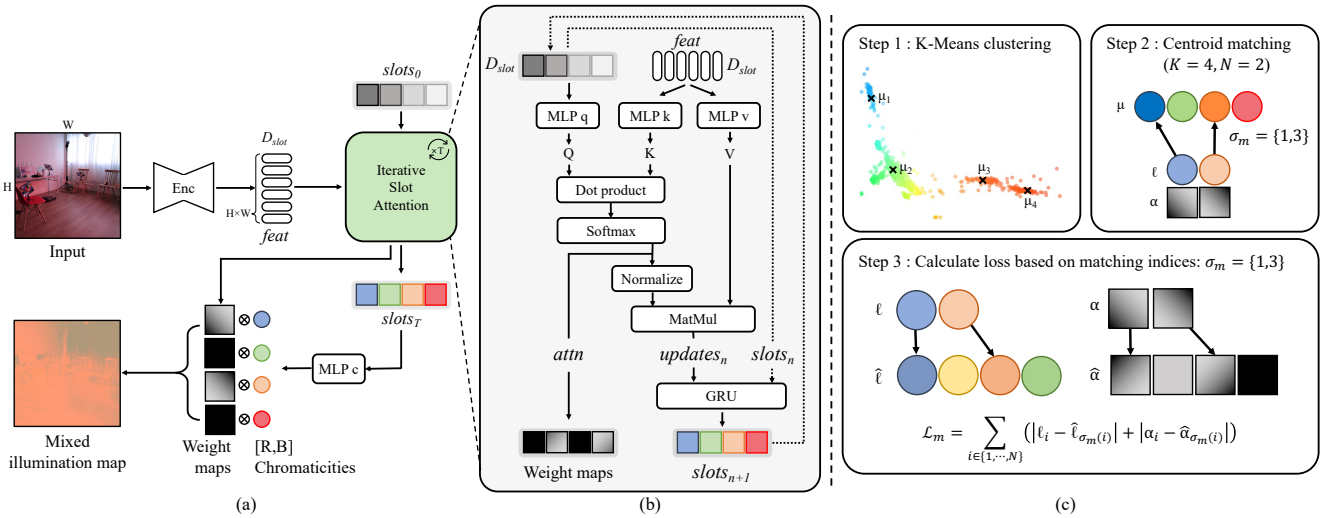

Figure 2. (a) Overview of our framework. Image feature is extracted from the input using an U-Net encoder. Next, the slot attention adaptively calibrates slot representation to be bound with illuminant chromaticity in each scene. Finally, the model fuses the chromaticity and the weight map to generate the mixed illumination map. (b) Detailed generation flow of weight maps and calibrated slots, where Q-Softmax denotes softmax application on the query dimension. (c) Illustration of the slot-wise loss using the centroid based Hungarian matching under $K = 4, N = 2$ assumption.

Image feature extraction. For a given raw image $\mathbf{I}$ with the resolution $H \times W$, the feature encoder E extracts a latent feature $feat$ with the same spatial resolution as the image and D_{slot} channels:

$$feat := E(\mathbf{I}) \in \mathbb{R}^{HW \times D_{slot}}, \tag{3}$$

where D_{slot} represents the dimension of slots.

Iterative slot calibration by slot attention. After extracting the image features, the iterative slot attention module (Fig. 2(b)) is applied to calibrate representations of the illuminant chromaticity. The details of the calibration process are as follows.

First, $slots_0 \in \mathbb{R}^{K \times D_{slot}}$ are initialized as learnable parameters where K indicates the number of slots. The slot attention module takes the image feature $feat$ and the initialized $slots$ as inputs to produce attention map $attn$:

$$\begin{aligned} attn_{i,j} &:= \frac{\exp(M_{i,j})}{\sum_l \exp(M_{i,l})}, \quad \text{where} \\ M &:= \frac{1}{\sqrt{D_{slot}}} k(feat) \cdot q(slots_n)^T \in \mathbb{R}^{HW \times K}, \end{aligned} \tag{4}$$

where k and q are MLPs for generating the key and query representations in D_{slot} dimension, and $slots_n$ represents the state of the slots in the n-th iteration.

Subsequently, the intermediate representation vectors, $updates$, are computed by aggregating the $values$ through spatially normalized attention map W:

$$\begin{aligned} updates &:= W^T \cdot v(feat) \in \mathbb{R}^{K \times D_{slot}}, \\ \text{where} \quad W_{i,j} &:= \frac{attn_{i,j}}{\sum_{l=1}^{N} attn_{l,j}}, \end{aligned} \tag{5}$$

where v is MLPs for generating value representation.

Finally, the calibrated $slots_{n+1}$ are refined by the GRU [12], which takes $updates$ as input and previous $slots_n$ as hidden state:

$$slots_{n+1} = GRU(slots_n, updates_n). \tag{6}$$

The process from Eq. (4) to Eq. (6) is repeated T times to generate final calibrated slots, $slots_T$.

Weight map & Illuminant chromaticity fusion. In our framework, we have carefully designed the output tensors of the slot-attention module, $slots$ and $attn$, to represent the chromaticity of illuminants $\boldsymbol{\ell}$ and the weight map α, respectively. Specifically, the $HW \times K$ shaped tensor $attn$, represents the pixel-wise similarity score between $feat$ and each $slot$, enabling its direct use as the set of K weight maps for each illuminant $(\hat{\alpha}_1 \dots \hat{\alpha}_K)$. Also, chromaticities for K illuminants $(\hat{\boldsymbol{\ell}}_1 \dots \hat{\boldsymbol{\ell}}_K)$ can be generated through passing calibrated $slots_T$ to chromaticity conversion MLPs c, where $c(slots_T)$ is a $K \times 2$ shaped tensor. Finally, we can fuse these two tensors to make mixed illumination map $\hat{\boldsymbol{\ell}}_{mixed}$ according to Eq. (2), by simply multiplicating them:

$$\hat{\boldsymbol{\ell}}_{mixed} = \sum_{k=1}^{K} \hat{\alpha}_k \hat{\boldsymbol{\ell}}_k = attn \cdot c(slots_T). \tag{7}$$

Although the above equation utilizes the final calibrated slots, $slots_T$, it is noteworthy that we can visualize the chromaticity $\hat{\boldsymbol{\ell}}_k$ and weight map $\hat{\alpha}_k$ for each iteration by employing the respective $slots_n$ of iteration n. Fig. 3 demonstrates how the generated illuminant chromaticity $\hat{\boldsymbol{\ell}}_k$ and weight map $\hat{\alpha}_k$ changes as $slots$ are itaratively calibrated.

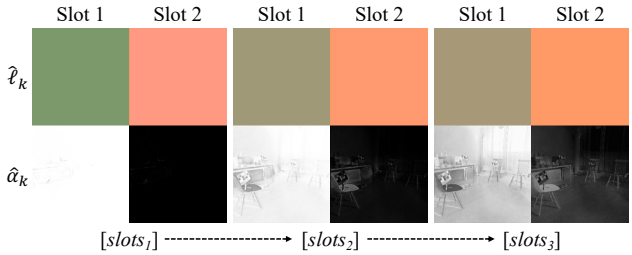

Figure 3. Slot calibration process. The chromaticity $\hat{\boldsymbol{\ell}}_k$ and weight map $\hat{\alpha}_k$ generated from each $slots_n$ are iteratively calibrated to their ground truth values.

3.3. Loss functions

AID framework is trained with two types of loss function: 1) mixed illumination loss, and 2) slot-wise loss, named as centroid-matching loss. The total training objective is to minimize the sum of these two loss functions.

Mixed illumination loss. Mixed illumination loss $\mathcal{L}_{mixed}$ is simply defined by L1 distance between the predicted mixed illumination map $\hat{\boldsymbol{\ell}}_{mixed}$ and the ground truth $\boldsymbol{\ell}_{mixed}$:

$$\mathcal{L}_{mixed} = \left| \boldsymbol{\ell}_{mixed} - \hat{\boldsymbol{\ell}}_{mixed} \right|. \tag{8}$$

Centroid-based matching loss. The mixed illumination loss alone does not provide a sufficient constraint that ensures our model activates the appropriate number of slots. Instead, it may result in the activation of either all slots or a random number of slots. As depicted in Fig. 2(a), the scene involves two illuminants, yet the model employs four slots to generate a mixed illumination map. Hence, it is necessary to strategically select and supervise the slots which are aligned with the ground-truth chromaticity and weight map. In this context, we design the loss term based on two assumptions that 1) each slot possesses its pre-defined cluster (color-gamut), and 2) activation of the slot should occur when the ground truth chromaticity lies within its cluster boundary. To this end, we propose the centroid matching loss and the calculation process of this loss is presented in Fig. 2(c).

First let us denote a pre-calculated set of centroids as $\mu = \{\mu_i\}_1^K$, obtained by applying K-means algorithm on the illuminant chromaticity distribution of the dataset. These centroids serve as the centerpoints of each illuminant chromaticity cluster, and in AID framework, each slot is responsible for representing one of these clusters. Next, we obtain a set of matched indices σ_m that minimizes the L1 cost between the matched centroid chromaticitiy μ and the ground truth chromaticity $\boldsymbol{\ell}$:

$$\sigma_m = \underset{\sigma}{argmin} \sum_{i}^{N} \left| \boldsymbol{\ell}_i - \mu_{\sigma(i)} \right|, \tag{9}$$

where N is the number of ground-truth illuminants in each scene and σ is one of the combinations of N elements from the set $\{1, \cdots, K\}$. Now we can define the loss term with respect to the chromaticity and weight map of each matched slots as follows:

$$\mathcal{L}_{centroid} = \sum_i \left(\left| \boldsymbol{\ell}_i - \hat{\boldsymbol{\ell}}_{\sigma_m(i)} \right| + \left| \alpha_i - \hat{\alpha}_{\sigma_m(i)} \right| \right), \tag{10}$$

where the centroid-matching loss $\mathcal{L}_{centroid}$ consists of both L1 loss for chromaticity and weight map of the matched slot indices. Here, the predicted weight map $\hat{\alpha}_k$ and the illuminant chromaticity $\hat{\boldsymbol{\ell}}_k$, are the k-th channel of $attn$ and $c(slots_T)$, as previously shown in Eq. (7).

		mean	median
(a)	AWB [1] †	9.54	8.19
	Patch CNN [8] †	4.82	4.24
	AngularGan [42] †	4.69	3.88
	TransCC [31] †	2.78	2.15
	LSMI-U [27]	2.31	1.89
	AID	**2.04**	**1.73**
	AID + MDL	**1.93**	**1.60**

		mean			median		
		gal.	son.	nik.	gal.	son.	nik.
(b)	LSMI-H [27]	3.06	3.21	2.99	2.54	2.89	2.61
	LSMI-U [27]	2.68	2.15	1.92	2.17	1.74	1.54
	AID	**1.66**	**1.66**	**1.71**	**1.41**	**1.35**	**1.34**

Table 2. Mean angular error (MAE) for the spatially varying illumination map on LSMI dataset: (a) all-in-one (cross-camera), (b) device-specific. † indicates that the results of [31] are referenced.

4. Experiments

4.1. Experimental setup

We validate the multi-illuminant WB performance of AID using two datasets: the LSMI dataset [27] captured with three cameras having different bit-depths and spectral sensitivities, and the Multi-Illumination In the Wild dataset (MIIW) [34], which is a versatile dataset covering various illumination-related tasks, including multi-illuminant WB.

We use seven slots ($K = 7$), 64 latent channels for D_{slot}, and calibrate slots three times ($T = 3$). For the evaluation, the green channel was inserted (G=1) to the mixed illumination map $\hat{\boldsymbol{\ell}}_{mixed}$, and the mean angular error (MAE) in degree was calculated with respect to the ground truth illumination map. For more detailed information, please refer to the supplementary materials.

4.2. Spatially varying white balance

Quantitative comparison. For the LSMI dataset, we evaluated AID under two settings to show the robustness of the proposed method: all-in-one cross-camera and device-specific setting. As shown in Table 2(a), AID achieves the state-of-the-art performance compared to all previous

LSMI [27]	Galaxy		Sony		Nikon	
	single	multi	single	multi	single	multi
LSMI-U [27]	2.85	2.55	1.92	2.34	1.49	2.30
AID	**1.19**	**2.03**	**1.01**	**2.16**	**1.11**	**2.26**

Table 3. Average MAE values obtained through experiments on the LSMI test set, distinguishing between single and multi-illuminant scenarios.

MIIW [34]	Single (1)		Multi (2,3)		Mixed (1,2,3)	
	mean	median	mean	median	mean	median
LSMI-U [27]	4.15	2.39	4.34	3.87	4.28	3.54
AID	**1.07**	**0.73**	**3.14**	**2.81**	**2.46**	**2.11**

Table 4. MAE values for predicting single-, multi-, and mixed-illuminant scenario using the MIIW test set.

models in LSMI dataset. Here, we would like to inform that LSMI-H and LSMI-U are the preceding state-of-the-art models introduced in the LSMI dataset. LSMI-H employs HDR-Net [18], while LSMI-U utilizes U-Net [38].

Moreover, as AID uses a concept of slots as an intermediate representation of the illumination, the model can be easily extended to multi-domain learning (MDL). We simply make the slot initialization different depending on the camera model (AID + MDL) and this slight modification brings additional 5% performance enhancement. Furthermore, in the camera-specific setting (Table 2 (b)), AID outperforms the LSMI baselines for all three cameras.

Since the LSMI dataset consists of one- to three-illuminant scenes, we also tested device-specific models with single and multi (two to three) illuminant subset, separately. As illustrated in Table 3, our framework consistently outperforms the LSMI baseline [27] across all devices in both single- and multi-illuminant settings. For the single-illuminant case, we could observe that only one slot is activated among 7 slots, and produces near-perfect global uniform illumination, resulting in a significant performance improvement over LSMI baseline.

We further demonstrate the robustness of our framework using another large-scale dataset, Multi-Illumination in the Wild [34]. Since no other algorithms have been applied to MIIW dataset previously, we select LSMI-U as our baseline. Table 4 demonstrates that AID outperforms LSMI-U, which had previously shown the best performance on the LSMI dataset, further highlighting the superior performance of AID.

Qualitative comparison. Fig. 4 illustrates the qualitative comparison between LSMI-U [27] and our method AID. For better visibility, we apply the following post-processing: 1) convert the result images in the top three rows and input images in the bottom two rows to the sRGB color space, and 2) scale down the green channel of the illumination maps in the bottom two rows. Our model generates more natural and ground truth-like WB results and illumination maps compared to the LSMI-U. We contribute the improvement of AID to the model design where the final illumination maps are generated under the condition of the physical image model Eq. (2), and also to the proposed loss function that matches the predictions to the proper ground truths.

NUS-8 [11]	Mean	Med.	Tri.	Best 25%	Worst 25%	G.M.
CCC [3]	2.38	1.48	1.69	0.45	5.85	1.73
AlexNet-FC4 [24]	2.12	1.53	1.67	0.48	4.78	1.66
FFCC [4]	1.99	1.31	1.43	**0.35**	4.75	1.44
CLCC [32]	1.84	1.31	1.42	0.41	4.2	1.42
AID	**1.57**	**1.03**	**1.16**	0.37	**3.67**	**1.21**

Table 5. Three-fold cross-validation result on NUS-8 dataset, with mean angular error in degrees.

We also provide the plots of chromaticity of the pixel-wise illumination predictions with the ground truths. Through the ground-truth illumination distributions of the top three rows, illustrated in red, it can be confirmed that each scene has a single, dual, or triple illuminant, respectively represented as a point, a line segment, and a triangle. It can be easily notified that the previous model generates unrefined predictions whereas our model performs well on reconstructing the actual distribution of the chromaticities of the illumination. For additional visualizations, including results related to the MIIW dataset, please refer to the supplementary material.

4.3. Generalization using single-illuminant DB

We also assessed the generalizability of our framework using the established single-illuminant white balance dataset, NUS-8 [11], which is a well-known benchmark widely used in the literature. NUS-8 dataset comprises 8 camera subsets, and we conducted three-fold cross-validation experiment for each camera. We measured the following metrics in the same way as previous studies [3, 4, 24]: mean, median, tri-mean, best 25%, worst 25%, and their geometric mean (G.M.). For the model configuration, we use $K = 5$, $T = 3$, and $D_{slot}, D_{attn} = 64$.

In Table 5, we report the angular error in degrees, along with the performance of recent works. The result shows that the proposed framework works robustly under both single- and multi-illuminant environments.

4.4. Fully decomposed multi-illuminant WB

Since our model generates fully decomposed illumination map, we can calculate the number of illuminants or the prediction accuracy of individual illuminant's chromaticity.

Count & chromaticity prediction result. We can also evaluate the accuracy of the predicted number of illuminants in the scene and the angular error of the chromaticities of each individual illuminant, using decomposed illu-

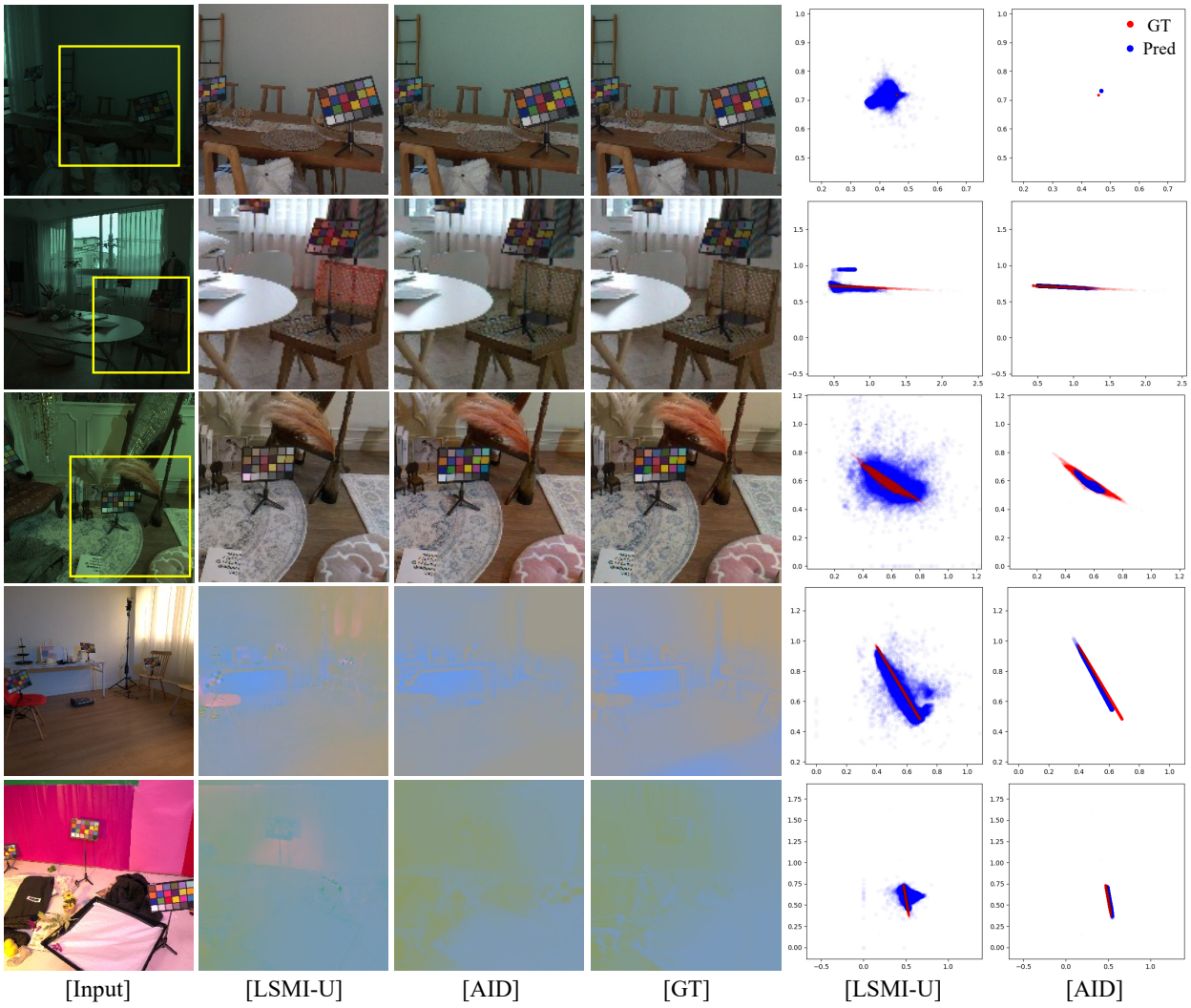

Figure 4. Qualitative comparison using LSMI test set. Top three rows show original raw image and corresponding WB results. The last two rows show the sRGB input images and corresponding illumination maps. The two rightmost columns demonstrate that our model, which infers illuminant-wise chromaticity and spatially mixes them, leads to more stable illumination plots compared to previous approaches. The x-axis and y-axis of the plot represent the ratio of the illumination value of the R and B channels to the value of the G channel.

	# of illum acc.	illuminant AE	
		mean	median
Galaxy	0.800	1.71	1.25
Sony	0.871	1.50	0.86
Nikon	0.813	1.84	1.20

Table 6. Additional validation metrics. We measured 1) the accuracy of predicted number of illuminants in the scene and 2) the angular error (AE) between predicted and the GT chromaticities.

mination map. The number of illuminants was measured by ignoring slots where the maximum value of the weight map component was below the threshold of 0.3. Angular error of illuminant chromaticity was measured between chromaticity vectors with matched indices σ_m and their corresponding GT vectors. Such additional information could be utilized as an additional metric for how well the model understands and accurately decomposes the multi-illuminant scene. Table 6 demonstrates that AID accurately predicts the chromaticity and the number of each illuminant. One thing to note is that it is impossible to measure this decomposition performance in previous works as ours is the first work to enable this illumination decomposition in multi-illuminant scenes.

Controllable multi-illuminant WB. Unlike previous multi-illuminant WB methods, AID can make fully-decomposed results with illuminant-wise chromaticities and weight maps. Therefore, we can leverage these decom-

Figure 5. Further applications of AID framework on LSMI test set examples. The separated weight map and the corresponding illuminant chromaticity (Decomp) allow for individual white balance to be applied to each light (WB Illum1,2), and for the chromaticity to be adjusted as desired (Illum manip). Full WB shows the results of applying white balance to all illuminants for reference. Gamma was adjusted for all images to increase visibility, and the G channel was scaled down for the decomposed illumination map visualization.

posed information to provide additional features like manipulating the chromaticity of each light or selective WB. Fig. 5 shows additional capabilities of AID framework.

4.5. Ablation study

As shown in the Table 7, we present three different ablation studies: centroid-matching loss ($\mathcal{L}_m$), the number of slots (K), and the number of iterations in the slot attention module (T). The first section of the table shows that the centroid-based matching loss helps the model to decompose the mixed illumination with the proper number of slots, as demonstrated by the illuminant number prediction accuracy (# acc.). Absence of $\mathcal{L}_{centroid}$ resulted in failure to effectively decompose mixed illumination using slots, as all slots were indiscriminately engaged to estimate the illumination, yielding an decomposition accuracy of 0.288. The efficacy of the centroid matching loss is more clearly demonstrated in Section C and Fig. 11 of the supplementary material. In addition, the second and the third section of the study reveals that the model performance can deliver different results depending on the number of slots (K) and the number of iterations in the slot attention module (T).

Among the combinations of K and T, we choose to use $K = 7$ and $T = 3$ combination by considering the accuracy and the computational cost. Ablation studies are conducted using the Galaxy camera subset of the LSMI dataset.

$\mathcal{L}_{mixed}$	$\mathcal{L}_{centroid}$	K	T	Mixed illum MAE		Illuminant	
				mean	median	# acc.	AE
✓		7	3	1.58	1.26	0.288	-
✓	✓	7	3	1.66	1.41	0.800	1.71
✓	✓	**5**	3	1.82	1.37	0.744	2.40
✓	✓	**7**	3	1.66	1.41	0.800	1.71
✓	✓	**9**	3	1.84	1.42	0.488	1.49
✓	✓	7	**2**	1.85	1.46	0.688	1.79
✓	✓	7	**3**	1.66	1.41	0.800	1.71
✓	✓	7	**4**	1.92	1.44	0.720	1.84

Table 7. Results of ablation studies on the centroid-matching loss ($\mathcal{L}_m$), the number of slots (K), and the number of iterations of GRU (T) using the Galaxy camera subset of the LSMI dataset.

5. Conclusion and discussion

In this paper, we introduced a framework called AID, designed to extract the chromaticity of individual illuminants along with their corresponding weights, while satisfying the linearity constraint of the Lambertian image model. To construct our model, we incorporated the slot attention module and applied the centroid-based matching loss, extending upon previous multi-illuminant white balance methods.

We demonstrated the effectiveness of AID through various experiments, and we believe this marks as a step towards more interpretable image enhancement, particularly in the context of white balancing. However, we acknowledge limitations in our proposed method, such as the requirement for presets regarding the number of clusters. Building model that can dynamically determine the number of clusters based on input images can be a promising path for future research.

Acknowledgement

This research was supported and funded by Artificial Intelligence Graduate School Program under Grant 2020-0-01361, Artificial Intelligence Innovation Hub under Grant 2021-0-02068

References

[1] Mahmoud Afifi, Brian Price, Scott Cohen, and Michael S Brown. When color constancy goes wrong: Correcting improperly white-balanced images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1535–1544, 2019.

[2] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Auto white-balance correction for mixed-illuminant scenes. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1210–1219, 2022.

[3] Jonathan T Barron. Convolutional color constancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 379–387, 2015.

[4] Jonathan T Barron and Yun-Ta Tsai. Fast fourier color constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–894, 2017.

[5] Shida Beigpour, Christian Riess, Joost Van De Weijer, and Elli Angelopoulou. Multi-illuminant estimation with conditional random fields. *IEEE Transactions on Image Processing (TIP)*, 23(1):83–96, 2013.

[6] Simone Bianco and Raimondo Schettini. Adaptive color constancy using faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(8):1505–1518, 2014.

[7] Simone Bianco, Claudio Cusano, and Raimondo Schettini. Color constancy using cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPR - Workshop)*, pages 81–89, 2015.

[8] Simone Bianco, Claudio Cusano, and Raimondo Schettini. Single and multiple illuminant estimation using convolutional neural networks. *IEEE Transactions on Image Processing (TIP)*, 26(9):4347–4362, 2017.

[9] Michael Bleier, Christian Riess, Shida Beigpour, Eva Eibenberger, Elli Angelopoulou, Tobias Tröger, and André Kaup. Color constancy and non-uniform illumination: Can existing algorithms work? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCV - Workshop)*, pages 774–781. IEEE, 2011.

[10] Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1): 1–26, 1980.

[11] Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *Journal of the Optical Society of America A (JOSA A)*, 31(5):1049–1058, 2014.

[12] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Advances in Neural Information Processing Systems Workshop (NeurIPS - Workshop)*, 2014.

[13] Florian Ciurea and Brian Funt. A large image database for color constancy research. In *Color and Imaging Conference*, pages 160–164. Society for Imaging Science and Technology, 2003.

[14] Martin Engelcke, Adam R Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. *arXiv preprint arXiv:1907.13052*, 2019.

[15] Graham D Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. In *Color and Imaging Conference*, pages 37–41. Society for Imaging Science and Technology, 2004.

[16] David A Forsyth. A novel algorithm for color constancy. *International Journal of Computer Vision (IJCV)*, 5(1):5–35, 1990.

[17] Peter Vincent Gehler, Carsten Rother, Andrew Blake, Tom Minka, and Toby Sharp. Bayesian color constancy revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[18] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[19] Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Generalized gamut mapping using image derivative structures for color constancy. *International Journal of Computer Vision (IJCV)*, 86(2-3):127–139, 2010.

[20] Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Improving color constancy by photometric edge weighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(5):918–929, 2011.

[21] Arjan Gijsenij, Rui Lu, and Theo Gevers. Color constancy for multiple light sources. *IEEE Transactions on Image Processing (TIP)*, 21(2):697–707, 2011.

[22] Xiangpeng Hao and Brian Funt. A multi-illuminant synthetic image test set. *Color Research & Application*, 45(6):1055–1066, 2020.

[23] Eugene Hsu, Tom Mertens, Sylvain Paris, Shai Avidan, and Frédo Durand. Light mixture estimation for spatially varying white balance. In *ACM SIGGRAPH*, pages 1–7, 2008.

[24] Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4085–4094, 2017.

[25] Zhuo Hui, Aswin C Sankaranarayanan, Kalyan Sunkavalli, and Sunil Hadap. White balance under mixed illumination using flash photography. In *International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2016.

[26] Zhuo Hui, Kalyan Sunkavalli, Sunil Hadap, and Aswin C Sankaranarayanan. Illuminant spectra-based source separation using flash photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6209–6218, 2018.

[27] Dongyoung Kim, Jinwoo Kim, Seonghyeon Nam, Dongwoo Lee, Yeonkyung Lee, Nahyup Kang, Hyong-Euk Lee, ByungIn Yoo, Jae-Joon Han, and Seon Joo Kim. Large scale multi-illuminant (lsmi) dataset for developing white balance algorithm under mixed illumination. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 2410–2419, 2021.

[28] Jinwoo Kim, Janghyuk Choi, Ho-Jin Choi, and Seon Joo Kim. Shepherding slots to objects: Towards stable and robust object-centric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19198–19207, 2023.

[29] Thomas Kipf, Gamaleldin F Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. *arXiv preprint arXiv:2111.12594*, 2021.

[30] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977.

[31] Shuwei Li, Jikai Wang, Michael S Brown, and Robby T Tan. Transcc: Transformer-based multiple illuminant color constancy using multitask learning. *arXiv preprint arXiv:2211.08772*, 2022.

[32] Yi-Chen Lo, Chia-Che Chang, Hsuan-Chao Chiu, Yu-Hao Huang, Chia-Ping Chen, Yu-Lin Chang, and Kevin Jou. Clcc: Contrastive learning for color constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8053–8063, 2021.

[33] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020.

[34] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4080–4089, 2019.

[35] Lawrence Mutimbu and Antonio Robles-Kelly. Multiple illuminant color estimation via statistical inference on factor graphs. *IEEE Transactions on Image Processing (TIP)*, 25 (11):5383–5396, 2016.

[36] Seoung Wug Oh and Seon Joo Kim. Approaching the computational color constancy as a classification problem through deep learning. *Pattern Recognition (PR)*, 61:405–416, 2017.

[37] Yanlin Qian, Ke Chen, Jarno Nikkanen, Joni-Kristian Kamarainen, and Jiri Matas. Recurrent color constancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5458–5466, 2017.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[39] Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd van Steenkiste, Filip Pavetić, Mario Lučić, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. *arXiv preprint arXiv:2206.06922*, 2022.

[40] Lilong Shi. Re-processed version of the gehler color constancy dataset of 568 images. *http://www. cs. sfu. ca/˜ color/data/*, 2000.

[41] Wu Shi, Chen Change Loy, and Xiaoou Tang. Deep specialized network for illuminant estimation. In *Proceedings of Proceedings of European Conference on Computer Vision (ECCV)*, pages 371–387. Springer, 2016.

[42] Oleksii Sidorov. Conditional gans for multi-illuminant color constancy: Revolution or yet another approach? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPR - Workshop)*, pages 0–0, 2019.

[43] Joost Van De Weijer, Theo Gevers, and Arjan Gijsenij. Edge-based color constancy. *IEEE Transactions on Image Processing (TIP)*, 16(9):2207–2214, 2007.

[44] Ruocheng Wang, Jiayuan Mao, Samuel J Gershman, and Jiajun Wu. Language-mediated, object-centric representation learning. *arXiv preprint arXiv:2012.15814*, 2020.

[45] Bolei Xu, Jingxin Liu, Xianxu Hou, Bozhi Liu, and Guoping Qiu. End-to-end illuminant estimation based on deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3616–3625, 2020.

[46] Huanglin Yu, Ke Chen, Kaiqi Wang, Yanlin Qian, Zhaoxiang Zhang, and Kui Jia. Cascading convolutional color constancy. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 12725–12732, 2020.

[47] Yi Zhou, Hui Zhang, Hana Lee, Shuyang Sun, Pingjun Li, Yangguang Zhu, ByungIn Yoo, Xiaojuan Qi, and Jae-Joon Han. Slot-vps: Object-centric representation learning for video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3093–3103, 2022.