# Generating Aligned Pseudo-Supervision from Non-Aligned Data for Image Restoration in Under-Display Camera

Ruicheng Feng[1] Chongyi Li[1] Huaijin Chen[2] Shuai Li[2] Jinwei Gu[3,4] Chen Change Loy[1]
[1]S-Lab, Nanyang Technological University [2]SenseBrain Technology
[3]The Chinese University of Hong Kong [4]Shanghai AI Laboratory
{ruicheng002, chongyi.li, ccloy}@ntu.edu.sg
{huaijin.chen, shuailizju}@gmail.com jwgu@cuhk.edu.hk

## Abstract

*Due to the difficulty in collecting large-scale and perfectly aligned paired training data for Under-Display Camera (UDC) image restoration, previous methods resort to monitor-based image systems or simulation-based methods, sacrificing the realness of the data and introducing domain gaps. In this work, we revisit the classic stereo setup for training data collection – capturing two images of the same scene with one UDC and one standard camera. The key idea is to "copy" details from a high-quality reference image and "paste" them on the UDC image. While being able to generate real training pairs, this setting is susceptible to spatial misalignment due to perspective and depth of field changes. The problem is further compounded by the large domain discrepancy between the UDC and normal images, which is unique to UDC restoration. In this paper, we mitigate the non-trivial domain discrepancy and spatial misalignment through a novel Transformer-based framework that generates well-aligned yet high-quality target data for the corresponding UDC input. This is made possible through two carefully designed components, namely, the Domain Alignment Module (DAM) and Geometric Alignment Module (GAM), which encourage robust and accurate discovery of correspondence between the UDC and normal views. Extensive experiments show that high-quality and well-aligned pseudo UDC training pairs are beneficial for training a robust restoration network. Code and the dataset are available at* https://github.com/jnjaby/AlignFormer.

## 1. Introduction

Under-Display Camera (UDC) is an imaging system with cameras placed underneath a display. It emerges as a promising solution for smartphone manufacturers to completely hide the selfie camera, providing a notch-free viewing experience on smartphones. However, the widespread commercial production of UDC is prevented by poor imaging quality caused by diffraction artifacts. Such artifacts are unique to UDC, caused by the gaps between display pixels that act as an aperture. As shown in Figure 1(a), typical diffraction artifacts entail flare, saturated blobs, blur, haze, and noise. The complex and diverse distortions make the reconstruction problem extremely challenging.

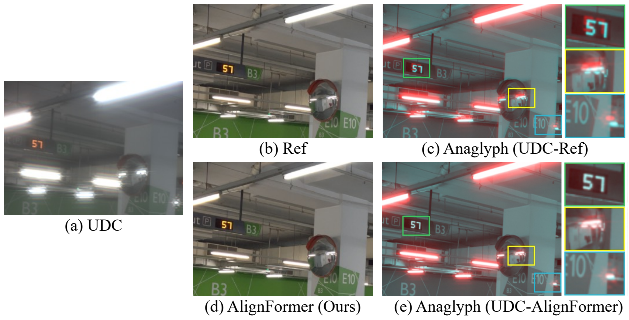

(a) UDC
(b) Ref
(c) Anaglyph (UDC-Ref)
(d) AlignFormer (Ours)
(e) Anaglyph (UDC-AlignFormer)

Figure 1. **Domain and geometric misalignment in UDC.** Stereo pairs (a) and (b) are captured by Under-Display Camera and high-end camera, respectively. The two images deviate significantly due to the color shift and severe degradation the UDC image. Anaglyph (c) illustrates the large spatial displacement between UDC and reference images despite a careful hardware setup and rough alignment. Our AlignFormer aligns the image pair and minimizes the parallax.

Training a deep network end-to-end for UDC restoration has been found challenging due to the need for a large-scale dataset of real-world degraded images and their high-quality counterparts. Existing methods [29, 52] build datasets with a monitor-based imaging system. As discussed in Feng *et al.* [3], such a paradigm is inherently limited by the dynamic range and spatial resolution of the monitor. To address the problem, Feng *et al.* [3] present a synthetic dataset grounded on the imaging formation model [3]. Both datasets exhibit degradation that deviates from the actual physical imaging process, leading to poor generalizability to diverse real-world test cases.

To circumvent the hurdle in collecting real paired data, we opt for an alternative setup, *i.e.*, to construct paired dataset with a *stereo* setting. Specifically, we capture two images of the same scene with one Under-Display Camera and one normal camera, denoted as UDC and Reference image, respectively. An example is shown in Figure 1(a-b) The key challenge lies in two aspects. i) **Domain discrepancy.** The different camera configurations inevitably give rise to variations in illuminance and severe color inconsistency, especially under the presence of color shift and severe diffraction artifacts in the UDC image. ii) **Geometric misalignment.** The contents in the UDC image and reference image are misaligned due to different focal lengths and field of views (FOV).

Due to the unique nature of UDC restoration, existing solutions are not effective in addressing the two aforementioned challenges. In particular, the low-level vision community has made attempts on this stereo setup for super-resolution [1], deblurring [32], and learnable ISP [9]. In addition, Contextual loss [25] and CoBi loss [48] are devised to alleviate mild spatial misalignment. As shown in our experiments, those methods are less stable and robust due to the difficulty of reliable matching when one image is severely distorted. In particular, the over-exposed regions caused by diffraction require strong pixel-wise supervision to enforce constraints during the training.

The key idea of our solution is to generate high-quality and well-aligned pseudo pairs from the non-aligned stereo data (UDC and reference) to enable end-to-end training of a deep network. The challenge lies in solving the domain and spatial misalignment so that the process resembles 'copying' details from the reference image selectively and then 'pasting' on the degraded image. To this end, we devise a simple yet effective Transformer-based framework, namely *AlignFormer*, with a Domain Alignment Module (DAM) and a Geometric Alignment Module (GAM). The DAM is inspired by AdaIN [7], aiming to mitigate the domain discrepancy between the UDC and reference images, allowing more robust and accurate correspondence matching in the subsequent stage. The GAM establishes accurate dense correspondences through incorporating geometric cues in attention. Specifically, GAM can flexibly work with any off-the-shelf pre-trained optical flow estimators to build pixel-wise correspondence between the UDC and reference images. The discovered correspondence then guides the sparse attention in our Transformer to search for the matching pixels accurately and effectively within local regions.

Figure 1(d-e) show that AlignFormer produces well-aligned image pairs. The results of AlignFormer can serve as pseudo ground-truth data and one can easily train an image restoration network end-to-end with common training settings, *i.e.*, using pixel losses such as $\mathcal{L}_1$ that assume exact spatial alignment, the perceptual loss [14], and the adversarial loss. Moreover, the constructed pseudo-paired dataset allows us to enjoy the merits of any advanced architectures of neural networks designed for image restoration problems. The generated data do not suffer from the limited dynamic range of spatial resolution as in previous monitor-based imaging systems. The data also experience a far lower domain gap than simulation-based approaches.

The main contributions are three-fold:

- We propose a data generation framework that is specifically designed for UDC. It presents a promising direction beyond previous monitor-based and simulation-based data collection approaches, leading to improved generalizability of UDC image restoration.
- Our AlignFormer properly integrates optical flow guidance into up-to-date Transformer architectures.
- Experimental results demonstrate significant progress in practical UDC image restoration in real-world scenarios.

## 2. Related Work

**UDC Image Restoration.** Very few works in the literature have investigated image restoration for UDC. Zhou *et al.* [52] and ECCV 2020 challenge [51] pioneered this line of works and inspired the follow-up studies [3, 18, 21, 27]. Yang *et al.* [45] proposed to redesign the pixel layouts for UDC display by optimizing the display patterns to improve the quality of restored image, which is orthogonal to our work. The dataset of the challenge [51, 52] is captured by a monitor-based imaging system. Such a system only induces incomplete diffraction artifacts due to the limited dynamic range of monitor [3]. Qi *et al.* [29] further explored the use of HDR monitor data. However, it is still inherently limited by the spatial resolution and contrast of the monitor. To remedy this issue, instead of capturing monitor-based image pairs, Feng *et al.* [3, 4] and Gao *et al.* [5] explored simulation pipelines for building synthetic dataset with real-captured point spread function (PSF) and imaging formation model. Despite well calibration and correction of PSF, models trained on synthetic dataset exhibit limited generalization capability for real-world images, especially for those with strong illuminations and flare regions. This is partially due to the domain shift between the mathematical model and physical imaging process in the real world, as there is no guarantee that the simulation pipeline can well approximate the complicated degradation in practical scenarios. Unlike previous works, we propose to collect real-world degraded-reference image pairs and produce high-quality images that contain the same content as degraded images as the pseudo target. The pseudo label allows us to enjoy the merits of advanced network architectures that are trained with pixel-wise losses.

**Dealing with Misaligned Paired Data.** Several studies have been devoted to capturing real-world image pairs using different cameras or camera configurations for other low-level tasks. Qu *et al*. [30] and Rim *et al*. [32] devised image acquisition systems with a beam splitter to collect paired data. Wang *et al*. [42] present dual-camera super resolution. Ignatov *et al*. [9] and Zhang *et al*. [49] collected image pairs and roughly aligned them via SIFT keypoints [23] and RANSAC algorithm [41]. The geometric alignment algorithm they adopted assumes image pairs can be aligned with a single homography, which is not the case where depth discrepancy exists in the scenes. Cai *et al*. [1] developed a pixel-wise image registration method to iteratively transform and adjust luminance. Nonetheless, this algorithm only works while the misalignment is mild. Contextual loss (CX) [25] was proposed to relax the constraints of spatial alignment and the loss works based on context and semantics. Inspired by CX [25], Zhang *et al*. [48] presented a contextual bilateral (CoBi) loss to prioritize local features and improve the matching quality. None of them consider image pairs exhibiting severe degradation and color inconsistency. Another line of studies focuses on transferring textures and details from the reference image, *e.g*., TTSR [46], $\mathcal{C}^2$-matching [12, 13]. Our framework is inspired by these prior studies. These algorithms only discover pixel correspondence through semantic similarity without using any geometric cues. Our approach differs in guiding the Transformer's attention with geometric prior and considering the additional domain alignment for addressing the significant gaps between the UDC and reference images.

## 3. Method

Given a UDC image $I_D \in \mathbb{R}^{H \times W \times 3}$ and a high-quality reference image $I_R \in \mathbb{R}^{H \times W \times 3}$ of the same scene, we aim to generate $I_P$ that possesses finer texture and details from $I_R$ and preserves the content of $I_D$,

$$I_P = \mathcal{T}(I_D, I_R; \Theta), \tag{1}$$

where $\mathcal{T}$ denotes the transformation. The whole process can be regarded intuitively as "copying" texture from reference images and "pasting" to the target image according to the semantic content of degraded images. The constructed pseudo pair $(I_D, I_P)$ is well-aligned and could serve as a training sample to provide better supervision for subsequent UDC image restoration networks $f_\theta$, given by

$$I_O = f_\theta(I_D), \tag{2}$$

where $I_O$ is the reconstructed clean image. The key challenge lies in aligning $I_R$ to $I_D$ in the presence of significant domain inconsistency. To overcome this, we propose a novel Transformer-based framework, *AlignFormer*, to mitigate both domain shift and spatial misalignment.

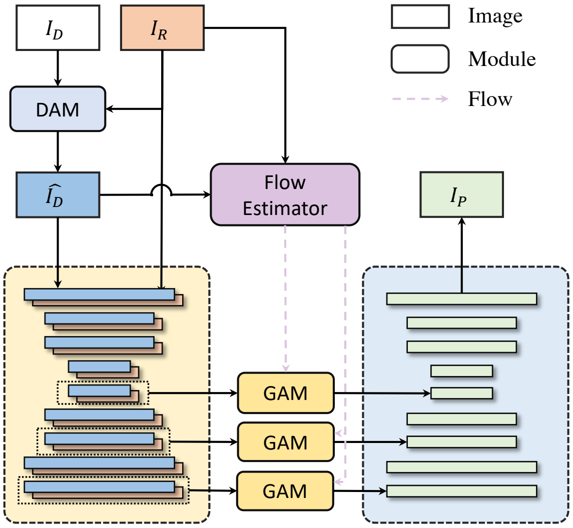


Figure 2. **Overview of the proposed AlignFormer.** We first mitigate domain discrepancy between UDC image $I_D$ and reference image $I_R$ via Domain Alignmain Module (DAM) to obtain $\hat{I_D}$, which are then gathered with $I_R$ and fed into two U-shape CNNs for feature extraction. Then the features at each scale are attended by the Geometric Alignment Transformer (GAM) to obtain the output features, which will be processed and fused in another U-Net to produce the pseudo image $I_P$.

### 3.1. AlignFormer

Inspired by Texformer [44], the overall architecture of AlignFormer $\mathcal{T}$ is illustrated in Fig. 2. It mainly consists of Domain Alignment Module (DAM), Geometric Alignment Module (GAM), flow estimator, and feature extractors. The DAM is carefully designed to modulate features towards reduced domain inconsistency, and consequently improve the accuracy of correspondence matching. On top of it, the GAM establishes accurate dense correspondences by incorporating geometric cues derived from any off-the-shelf optical flow estimators in attention. This enables sparse attention in our Transformer to search for the matching pixels accurately and effectively within local regions.

**Domain Alignment Module (DAM).** From our experiments, we found that the performance of attention is highly susceptible to domain shift and severe degradations. Inspired by the recent success of style-based architecture [15, 16], we propose DAM to mitigate domain inconsistency between the UDC and reference images. The structure of DAM is shown in Fig. 3, consisting of two sub-nets, a guidance net and a matching net.

The guidance net generates a conditional vector as guidance for the matching net by extracting and exploring feature statistics, *i.e*., domain information, from the reference image $I_R$. To obtain the guidance vector, denoted as $\boldsymbol{s} \in \mathbb{R}^d$, we compose a stack of convolution layers, followed by a global average pooling layer. The guidance vector can serve as the condition and deliver holistic information from

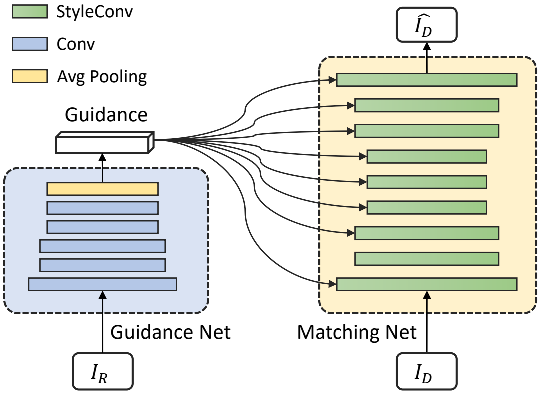


Figure 3. **The structure of domain alignment module.** This module comprises a guidance net and a matching net. The guidance vector generated by the guidance net is used for style modulation via StyleConv in the matching net. Such designs help to imitate the color and illuminance of reference image, while preserving spatial information of UDC image.

the reference image to the matching net.

The matching net is designed to transfer domain information, *e.g.*, color, illuminance, contrast, to the degraded UDC image. To leverage the style condition $\boldsymbol{s} \in \mathbb{R}^d$ and match the feature, we utilize StyleConv [15], consisting of a Conv layer, affine transformation, and AdaIN [7]. Specifically, let $\mathbf{A} \in \mathbb{R}^{2d \times d}$ and $\mathbf{b} \in \mathbb{R}^{2d}$ be the layer-wise learnable affine transformation applied to $\boldsymbol{s}$. For each AdaIN layer, we can define the style input as $\boldsymbol{y} = [\boldsymbol{y}_s, \boldsymbol{y}_b] = \mathbf{A}\boldsymbol{s} + \boldsymbol{b}$. Given the output feature of Conv layer $\boldsymbol{x}$ as the content input and style input $\boldsymbol{y}$, AdaIN performs

$$\text{AdaIN}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y}_s \frac{\boldsymbol{x} - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})} + \boldsymbol{y}_b. \tag{3}$$

The output of DAM, denoted as $\hat{I_D}$, will exhibit a closer style to the $I_R$, and will be used for the subsequent GAM. Note that the style transfer is performed from $I_R$ to $I_D$, and not otherwise, as the latter will hamper the correspondence discovery due to poorer quality of image pairs.

**Geometric Alignment Module (GAM).** The conventional formulation in Transformers [2, 22, 40] is not well-suited for our task. In particular, the vanilla Transformers [2, 40] exhibit the superiority in exploiting semantic similarity and capturing global contextual information. Nonetheless, their global attention densely attends to all key components, causing a diversion of attention to irrelevant and redundant elements. Swin Transformer [22], on the other hand, assumes attention within local regions, which is better suited for our task. However, it is not readily designed to accept geometric cues (*e.g.*, optical flow, patch matching) as guidance, which is crucial in our task considering local rigidity assumption under the stereo setting. Hence, to discover correspondences between the UDC input $I_D$ and reference image $I_R$, a specialized attention mechanism is required.

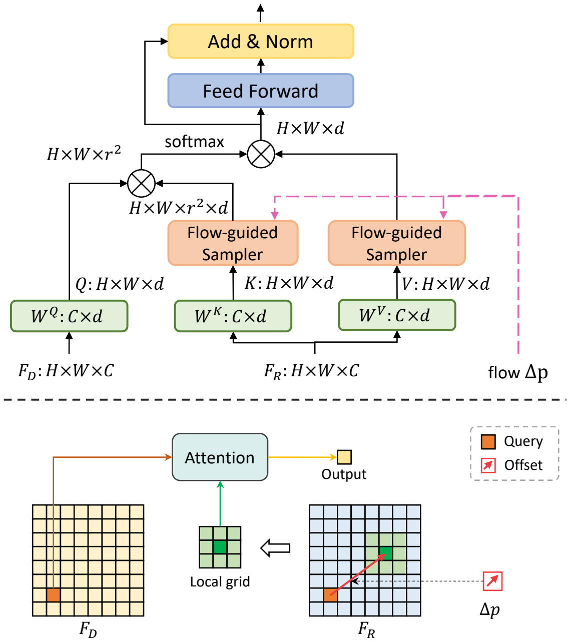


Figure 4. Top: The structure of geometric alignment module. Bottom: the schematic of flow-guided sampler.

To this end, we develop GAM to exploit the pixel correspondences by incorporating geometric cues in attention. Specifically, we use an off-the-shelf flow estimator[1] as the guidance to sample features from the vicinity in the reference image. We introduce a flow estimator prior in the conventional attention as it can exploit the geometric prior to facilitate the subsequent attention mechanism. Otherwise, it becomes knotty given the spatial misalignment between $\hat{I_D}$ and $I_R$. As shown in Figure 3.1, we reformulate the basic attention unit. It consists of attention with a flow-guided sampler, a layer normalization (LN), and a multi-layer perceptron (MLP).

To take advantage of the geometric cues, the *key* sampling is directed by the geometric information predicted by the pre-trained optical flow estimator. The dense offset map at each position $\mathbf{p} = (x, y)$ in $\hat{I_D}$ is mapped to its estimated correspondence in $I_R$: $\mathbf{p}' = (x + u, y + v)$, where the flow offset is recovered by a network $\psi_{flow}$

$$\Delta\mathbf{p} = (u, v) = \psi_{flow}(\hat{I_D}, I_R)(x, y). \tag{4}$$

Then a flow sampler defines the local grid around $\mathbf{p}'$

$$\mathcal{N}(\mathbf{p}')_r = \{\mathbf{p}' + \mathbf{d} \mid \mathbf{d} \in \mathbb{Z}^2, \|\mathbf{d}\|_1 \leq r\}. \tag{5}$$

The offsets within a radius of $r$ units are used to sample key and value elements. The flow sampler samples sub-pixel locations of real values from $\mathcal{N}(\mathbf{p}')_r$ via interpolation.

We denote $F_D \in \mathbb{R}^{H \times W \times C}$ and $F_R \in \mathbb{R}^{H \times W \times C}$ the feature extracted from $\hat{I_D}$, $I_R$, where $H$ and $W$ indicate

[1] We use RAFT [39] but other methods are applicable.

height and width of the feature map, and $C$ is the feature dimension. Given a linear projected query vector $\boldsymbol{q}_{x,y} = F_D W^Q$ at coordinate $\mathbf{p} = (x, y)$ of $F_D$, the flow-guided attention can be written as:

$$f_{attn}(\boldsymbol{q}_{x,y}, F_R) = \sum_{(i,j)\in\mathcal{N}(\mathbf{p}')_r} \text{sim}(\boldsymbol{q}_{x,y}, \boldsymbol{k}_{i,j})\boldsymbol{v}_{i,j}, \quad (6)$$

where $\boldsymbol{k}_{i,j} = F_R W^K$ and $\boldsymbol{v}_{i,j} = F_R W^V$ represent the projected vectors sampled from $F_R$ by the flow sampler. Here $W^Q$, $W^K$, $W^V \in \mathbb{R}^{C\times d}$ are the respective learnable linear projection for query, key, and value elements, where $d$ is the dimension of the projected vector. The attention score $\text{sim}(\boldsymbol{q}_{x,y}, \boldsymbol{k}_{i,j})$ is the scaled dot-product attention followed by softmax function, formulated as

$$\text{sim}(\boldsymbol{q}_{x,y}, \boldsymbol{k}_{i,j}) = \text{softmax}(\frac{\boldsymbol{q}_{x,y}^\top \boldsymbol{k}_{i,j}}{\sqrt{d}}). \quad (7)$$

Hence, the final output attended features are computed as

$$\begin{aligned} \boldsymbol{f}_{x,y} &= f_{attn}(\boldsymbol{q}_{x,y}), \\ \boldsymbol{z}_{x,y} &= f_{\text{MLP}}(f_{\text{LN}}(\boldsymbol{f}_{x,y})) + \boldsymbol{f}_{x,y}. \end{aligned} \quad (8)$$

Here $f_{\text{LN}}$ is the LayerNorm layer.

### 3.2. Learning Objectives

The training of AlignFormer requires an objective function that does not forcefully match each spatial position as the problem lacks exact spatially aligned supervision. The Contextual Loss (CX) [25] is a viable choice as it treats features of images as a set and measures the similarity between images, ignoring the spatial positions of the features. This property enables us to compare images that are spatially deformed.

Given two images $x$ and $y$, CX loss aims to minimize the summed distance of all matched feature pairs, formulated as

$$\mathcal{L}_{CX}(x, y) = \frac{1}{N}\sum_j \min_i \mathbb{D}(\phi(x)_j, \phi(y)_i), \quad (9)$$

where $\phi(x)_j$ and $\phi(y)_i$ are the $j$-th point of $\phi(x)$ and $i$-th point of $\phi(y)$, respectively. $\phi(x)$ denotes feature maps of $x$ extracted from the VGG network $\phi$, and $\mathbb{D}$ is some distance measure. Based on context and semantics, the CX loss transfers the style of an image to another by comparing regions with similar semantic meaning in both images. The insensitivity of CX loss for misaligned data is well-suitable for moderating the domain gap between UDC images and high-quality reference images used in our pipeline.

**Learning Objective for DAM.** We first train DAM to mitigate the domain shift. The loss term for DAM is given by

$$\mathcal{L}_{DAM} = \mathcal{L}_{CX}(\hat{I_D}, I_R), \quad (10)$$

where $\hat{I_D}$ denotes the output of DAM, $I_D$ and $I_R$ the degraded and reference image, respectively. We use the pre-trained VGG-19 [35] and select "conv4_4" as deep features.

**Learning Objective for AlignFormer.** After training DAM, we integrate the DAM and the pre-trained RAFT [39] into the AlignFormer. Note that both DAM and the optical flow estimator are fixed during training AlignFormer. Similarly, the AlignFormer is trained with a domain loss:

$$\mathcal{L}_{Align} = \mathcal{L}_{CX}(I_P, I_R), \quad (11)$$

where $I_P$ is the output of AlignFormer. The reference image $I_R$ serves as the domain supervision.

### 3.3. Image Restoration Network

Although AlignFormer can achieve good results with a UDC image and a reference image as input, a restoration network is still indispensable. This is because the reference image is not available during inference and some regions in AlignFormer's results have to be masked out due to the effect of occlusion between the UDC image and the reference image. Therefore, after getting the aligned pseudo image pairs ($I_D$, $I_P$), we devise a baseline network targeting at restoring both global information (*e.g.*, brightness and color correction), and local information (*e.g.*, flare removal, texture enhancement, denoising) to evaluate the effectiveness of our AlignFormer. While a tailored image restoration backbone UNet [33] could inherently enhance local details at different scales, it can hardly alter the image globally. To address this, we adopt a standard U-Net architecture with Pyramid Pooling Module (PPM) [50], namely *PPM-UNet*. The role of PPM is to incorporate global prior into the networks to mitigate the color inconsistency between UDC and generated images. Due to limited space, we leave the details of the restoration network to the supplementary material.

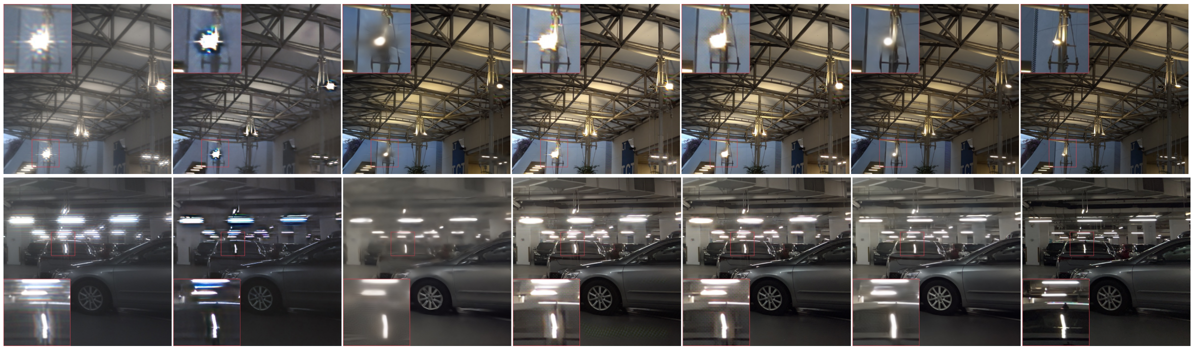
**Learning Objective for PPM-UNet.** Following common practice [43, 49], we train PPM-UNet with a combination of losses, including $\mathcal{L}_1$ loss, VGG-based perceptual loss [14] $\mathcal{L}_{VGG}$, and GAN loss, which can be defined by

$$\begin{aligned} \mathcal{L}_{rec} =& \lambda_1\|M \odot (I_P - I_O)\|_1 \\ &+\lambda_{VGG}\|\phi(M \odot I_P) - \phi(M \odot I_O)\|_1 \\ &+\lambda_{GAN}\mathcal{L}_{GAN}, \end{aligned} \quad (12)$$

where $\odot$ denotes element-wise multiplication, $\|\cdot\|_1$ is $\ell_1$-norm, $\phi$ is the pre-trained VGG [35] network, and $M$ is the valid mask indicating the non-occluded regions of optical flow. To avoid deteriorating networks due to inaccurate deformations over occluded regions, we mask out those pixels invisible in the reference image. The occlusion detection is derived from forward-backward consistency assumption [38]. To further improve the visual quality, we also add adversarial loss based on conditional PatchGAN [10]. Please refer to the supplementary material for details.

Table 1. **Comparison between different datasets on the baseline network.** We train the PPM-UNet using different training sets and evaluate their performance on the test set. Note that for non-aligned data, we also use CX [25] loss and CoBi [48] loss to tackle with misalignment. The best and runner up results are highlighted in **bold** and underlined, respectively.

| Training sets | Aligned | Loss | *Aligned Ref* |  |  | *Original Ref* |  | *Non Ref* |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  |  |  | PSNR↑ | SSIM↑ | LPIPS↓ | CD (L / a / b)↑ | SIFID $_{(\times 10^{-5})}$↓ | NIQE↓ | MUSIQ↑ | NRQM↑ |
| Synthetic [3] | ✓ | $\mathcal{L}_{rec}$ | 19.03 | 0.7808 | 0.3513 | 0.67 / 0.40 / 0.27 | 6.3341 | 6.4706 | 34.0738 | 4.8640 |
| Real |  | $\mathcal{L}_{rec}$ | 19.04 | 0.8187 | 0.1979 | 0.93 / 0.46 / 0.47 | 2.2157 | 7.5641 | 53.1251 | 5.8763 |
| Real |  | $\mathcal{L}_{CX}$ | 20.85 | 0.8198 | 0.1524 | 0.93 / 0.25 / 0.40 | 1.1508 | 9.7242 | 48.8314 | 5.9143 |
| Real |  | $\mathcal{L}_{CoBi}$ | 21.57 | 0.8319 | 0.1252 | 0.93 / 0.30 / 0.41 | 1.0385 | 8.9563 | 50.8363 | 6.0025 |
| Real | AlignFormer | $\mathcal{L}_{rec}$ | **22.95** | **0.8581** | **0.1236** | **0.94 / 0.48 / 0.47** | **0.9735** | **6.2816** | **56.3314** | **6.4839** |



UDC | *Synthetic* + $\mathcal{L}_{rec}$ | *Real* + $\mathcal{L}_{rec}$ | *Real* + $\mathcal{L}_{CX}$ | *Real* + $\mathcal{L}_{CoBi}$ | **Ours*** | Ref

Figure 5. **Visual comparison between different datasets on the baseline network.** * indicates results of "AlignFormer + PPM-UNet"

## 4. Experiments

**Data Collection and Pre-Processing.** To build the degraded-reference image pairs, we construct a stereo smartphone array - ZTE Axon 20 with selfie under-display camera, and iPhone 13 Pro rear camera, which are physically aligned as much as possible (see supplementary material for details). To eliminate the effects of built-in ISPs, both UDC and Ref images are extracted from raw dump of data with minimal processing (demosaic and gamma correction) and converted into sRGB domain. In total, we collect 330 image pairs covering both indoor and outdoor scenes.

Given a pair of images, we first roughly align them using a homography transformation estimated via RANSAC algorithm [41] as it is robust to photometric misalignment. Even after alignment with homography, there still exists mild misalignment between the image pair. This is because the geometric transformation is applied globally with the assumption that all points are located at the same plane in the world, which does not hold true where contents are at different depths in the scene. The remaining displacement in the pair will be further resolved by our AlignFormer.

**Implementation.** We split 330 image pairs into 274 pairs for training and 56 for testing. For each pair of full-resolution images, we crop them into $512 \times 512$ patches for training. More example training patches can be found in the supplement. We initialize all networks with Kaiming Normal [6] and train them using Adam optimizer [19] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\theta = 10^{-8}$, and the mini-batch size is set to 8 for all the experiments. $\lambda_1$, $\lambda_{VGG}$, and $\lambda_{GAN}$ are set to $10^{-2}$, 1, $5 \times 10^{-3}$, respectively. The learning rate is decayed by half at $250k$ and $300k$ iterations with a multi-step schedule. We implement our models with PyTorch [28] and train them using two NVIDIA V100 GPU cards.

### 4.1. Comparisons

**Evaluation Protocol.** In the absence of ground truths with spot-on alignment, it is challenging to evaluate the quality of generated images. To quantify the performance, we compute three common metrics (*e.g.*, PSNR, SSIM, and LPIPS [47]) on the Pseudo GT $I_P$, namely *Aligned Ref*. We additionally adopt two sets of evaluations for a comprehensive and accurate comparison. To quantify the realism of the generated images and how well they capture the domain information (*e.g.*, color, illuminance, contrast) of the reference images (*Original Ref*), we use two measures: CD - Color Distribution [10], and SIFID [34] - Single Image FID. Specifically, color distribution is to calculate the histogram intersection of marginal color distributions between our results and reference images in Lab color space. SIFID is to measure the internal statistics of patch distributions of single image. In addition, we adopt two non-reference (*Non Ref*) metrics MUSIQ [17], NIQE [26], and NRQM [24] to supplement the quantitative comparision.

**Dataset Comparison.** Before benchmarking existing approaches, we first conduct experiments to compare the performance of our datasets with other synthetic datasets [3]. Specifically, we train the image restoration network, PPM-UNet, of the same architecture, using several possible combinations of different training sets and learning objectives. Then, we evaluate their performance on the real test sets and summarize in Table 1. The column "Aligned" indicates

Table 2. **Benchmark of state-of-the-art UDC image restoration methods on real dataset.** "*" indicates checkpoint models released by the original paper. Others are retrained with our dataset. The best and runner up results are highlighted in **bold** and underlined, respectively.

| **Method** | *Aligned Ref* | | | *Original Ref* | | *Non Ref* | | |
|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | CD (L / a / b)↑ | SIFID $(\times 10^{-5})$↓ | NIQE↓ | MUSIQ↑ | NRQM↑ |
| DISCNet* [3] | 19.25 | 0.7574 | 0.3836 | 0.72 / 0.43 / 0.27 | 5.6081 | 6.0847 | 28.0298 | 5.9767 |
| BNUDC* [20] | 19.42 | 0.7502 | 0.3574 | 0.73 / 0.40 / 0.28 | 4.9629 | 6.9921 | 30.3077 | 6.5138 |
| MUNIT [8] | 20.61 | 0.8101 | 0.3258 | 0.93 / 0.44 / 0.41 | 3.1513 | 7.1354 | 36.2816 | 5.5094 |
| TSIT [11] | 19.09 | 0.7755 | 0.2167 | 0.69 / 0.41 / 0.42 | 2.1232 | **4.6044** | 48.8685 | **6.6491** |
| AlignFormer + DISCNet [3] | 21.66 | **0.8582** | 0.1452 | 0.90 / 0.47 / 0.46 | 1.1623 | 6.7404 | **56.3539** | 6.0486 |
| AlignFormer + BNUDC [20] | 20.43 | 0.8430 | 0.1589 | 0.87 / 0.45 / 0.45 | 2.4669 | 6.5180 | 55.1308 | 5.9677 |
| AlignFormer + PPM-UNet (Ours) | **22.95** | 0.8581 | **0.1236** | **0.94 / 0.48 / 0.47** | **0.9735** | 6.2816 | 56.3314 | 6.4839 |

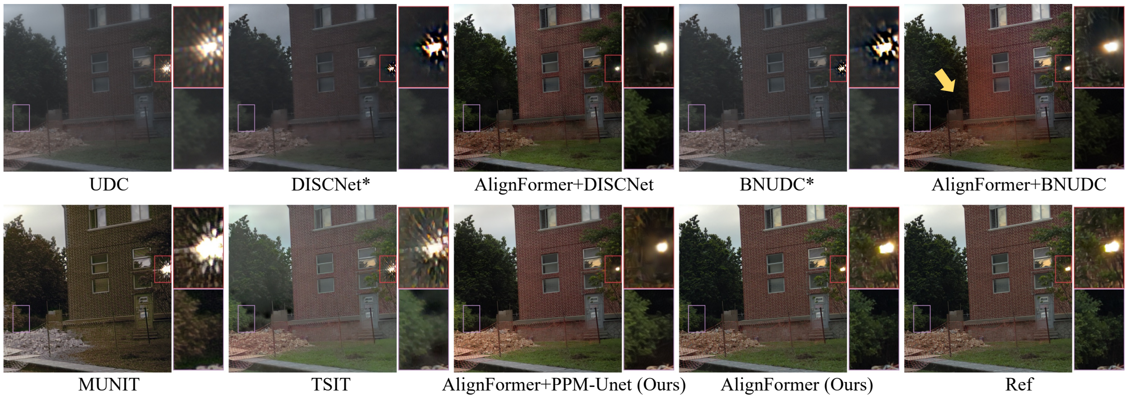


Figure 6. **Visual comparison of benchmarks and our method.** "AlignFormer+" indicates methods trained on pseudo image pairs generated by AlignFormer. Methods marked with * are pre-trained on synthetic dataset.

whether the image pairs are aligned. The table shows that models trained on the real dataset (even without any alignment) achieve higher performance on our test sets compared to the synthetic dataset in general, which proves that the synthetic dataset is not realistic enough to cover real-world flare images. Due to the considerable domain gap between synthetic and real data, the model trained on the synthetic dataset (1st row) significantly deteriorates performance on the real data, which validates the necessity of capturing real-world data. Besides, our baseline network trained on pseudo pairs (last row) demonstrates superior performance against all other methods. It also achieves the highest PNSR and SSIM scores when compared with the models trained using the misaligned UDC-Reference image pairs regardless of the use of reconstruction loss, CX loss, and CoBi loss. The quantitative comparisons suggest that the well-aligned and high-quality image pairs are indispensable for the training of UDC image restoration network, which can be achieved by our AlignFormer. We also show the visual comparisons in Figure 5. We can see that the model trained using the synthetic image pairs still cannot produce a satisfactory result, especially the overall perceptual quality and regions around the saturation. Moreover, pixel-wise supervision on misaligned image pairs (3rd column) incentivizes blurry results and severe artifacts. Although CX loss and CoBi loss are devised to alleviate inaccurate alignment, they fail to produce flare-free images and exhibit a different style (*e.g.*, brightness) compared to reference images. This stems from the absence of strong spatial constraints on flare regions and they are usually performed in feature space. In contrast, our AlignFormer effectively copies the fine details and textures of the reference image and pastes them back while maintaining alignment with the UDC image. Such high-quality and well-aligned image pairs, in turn, leading to the visually pleasing result when the model is trained using them, as the "**Ours**" shown in Figure 5.

**Bechmarking State-of-the-art Methods.** There exist several categories of approaches that could handle this problem. The direct rival to our method is training models on synthetic datasets and testing on real data, including DISCNet [3] and BNUDC [20]. Besides the officially released pre-trained models, we also include these networks trained with our dataset and training strategy. Image-to-image translations offer another avenue to solve this task. It remains challenging to build mappings between complex domains using existing style transfer algorithms, and they are usually dominated by global image distribution and overlook the detailed local structures (*e.g.*, flare, blur, noise). We include two representative image translation methods, MUNIT [8] and TSIT [11] for comparison.

Table 2 summarizes the benchmark. Pre-trained models (*i.e.*, DISCNet and BNUDC) achieve relatively low PSNR, SSIM and LPIPS values compared to *Aligned Ref*, and it shows poorer performance on CD and SIFID, indicating domain discrepancy against the reference images. In contrast, the same models retrained on our dataset (5th and 6th row) demonstrate noticeable improvement over these two sets of metrics. Image translation approaches only capture statistics over the distribution of the whole dataset, and thus neglect local details, leading to a performance drop on the

Table 3. **Ablation experiments on AlignFormer.** We report PCK (%)↑ with $\alpha = 0.01, 0.03, 0.10$. Our settings are marked in gray.

(a) Ablation study on alignment method.

| Alignment | Ref | Cai *et al.* [1] | RAFT [39] | AlignFormer |
|---|---|---|---|---|
| $\alpha = 0.01$ | 21.77 | 26.15 | 56.14 | 58.75 |
| $\alpha = 0.03$ | 62.02 | 67.02 | 94.79 | 95.08 |
| $\alpha = 0.10$ | 85.38 | 79.61 | 98.74 | 99.93 |

(b) Ablation study on optical flow estimator.

| Estimator | Zero | SPyNet [31] | PWC-Net [36] | RAFT [39] |
|---|---|---|---|---|
| $\alpha = 0.01$ | 9.61 | 31.31 | 33.66 | 58.75 |
| $\alpha = 0.03$ | 26.86 | 57.45 | 93.22 | 95.08 |
| $\alpha = 0.10$ | 36.30 | 67.60 | 98.49 | 99.93 |

(c) Ablation study on radius.

| Radius | 0 | 1 | 2 |
|---|---|---|---|
| $\alpha = 0.01$ | 56.18 | 58.23 | 58.75 |
| $\alpha = 0.03$ | 94.97 | 95.07 | 95.08 |
| $\alpha = 0.10$ | 98.71 | 98.72 | 99.93 |

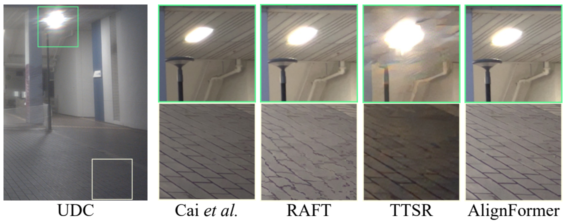


Figure 7. **Ablation study on alignment method.**

Table 4. **Ablation experiments on DAM and PPM.**

(a) Ablation study on DAM.

| PCK (%)↑ | w/o DAM | w/ DAM |
|---|---|---|
| $\alpha = 0.01$ | 56.15 | 58.75 |
| $\alpha = 0.03$ | 94.70 | 95.08 |
| $\alpha = 0.10$ | 98.66 | 99.93 |

(b) Ablation study on PPM.

| | w/o PPM | w/ PPM |
|---|---|---|
| PSNR↑ | 22.68 | 22.95 |
| SSIM↑ | 0.8670 | 0.8581 |
| LPIPS↓ | 0.1328 | 0.1236 |

test set. The PPM-UNet trained with image pairs generated from AlignFormer achieves comparable or the best performance on all sets of measurements. These results show the effectiveness of our AlignFormer and PPM-UNet.

Figure 6 qualitatively compares the benchmarks in Table 2. All pre-models fail to suppress flare and blur, and they cannot restore plausible color and contrast. On the other hand, the models trained with our datasets show better-restored results. The results of "AlignFormer+DISCNet" and "AlignFormer+BNUDC" accurately remove flare. However, the result of "AlignFormer+BNUDC" suffers from regional inconsistency (See yellow arrow). Image translation approaches cannot accurately recover the light source regions, and produce considerable noise and artifacts. Among these benchmarks, our results have sharper details with a similar domain style.

## 4.2. Ablation Study

**Effectiveness of Alignment Method.** As there is no ground-truth correspondence, it is non-trivial to quantify how well the pseudo GT is aligned to the UDC image, especially when severe domain inconsistency exists. Thus, we indirectly measure the displacement error with LoFTR [37] serving as a keypoint matcher. Given a set of matched keypoints for both images, PCK measures the percentage of correct keypoints transferred to another image, which lie within a certain radius of the same coordinates (ideally 0 if two images are well-aligned). Please refer to the supplementary material for details. Suppose $d$ is the displacement of a pair of matched keypoint, the keypoint pair is correctly aligned when $d < \alpha \times \max(H, W)$, where $\alpha$ is the threshold and $H$, $W$ represent the height and width of the image. Table 3a shows the accuracy of alignment among warping-based methods (RAFT [39], Cai *et al.* [1]). As presented, our AlignFormer more accurately align to the corresponding UDC images than the compared methods.

Besides visualizing the results of warping-based methods, we also include the result of TTSR [46] as reference in Figure 7. Flow-based alignment yields distorted structures, while Cai *et al.* [1] introduces illuminance change and blur. TTSR cannot recover high-quality details around the light region. Our method can alleviate spatial misalignment well.

**Optical Flow Guidance.** We incorporate three pre-trained optical flow estimators: SPyNet [31], PWC-Net [36], and RAFT [39] into our AlignFormer. The results in Table 3b show that the performance can be improved with better optical flow estimation. The model with higher accuracy (RAFT) exhibits less displacement on the resulting image. Removing flow guidance, denoted as *Zero*, remarkably deteriorates the performance.

**Radius of Local Grid.** As shown in Table 3c, we change the radius $r$, which specifies the size of the local grid used to search neighboring samples in the flow-guided sampler. When the radius is 0, the feature is retrieved at one single point given by the flow offset, which can be regarded as a warping strategy in the feature space. It can be observed that even with a radius 0, we can still get rough information from the optical flow.

**Effectiveness of DAM.** Table 4a shows that the domain alignment module drives more accurate alignment for subsequent GAM and facilitates more robust attention compared to that without DAM.

**Effectiveness of PPM.** The PPM layers propagate global prior into the network and stabilize training and suppress artifacts in UDC image restoration. In Table 4b, we remove Pyramid Pooling Module (w/o PPM) and the performance drops noticeably.

## 5. Conclusions

We have presented AlignFormer for generating high-quality and well-aligned pseudo UDC pairs. The key insight of our pipeline is to exploit the pixel correspondence by both semantic and geometric cues embedded into a Transformer-based structure. Apart from the novel designs, we also contribute a new dataset to support UDC image restoration networks training, a step towards solving UDC image restoration in the real world.

**Acknowledgement.** This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, a Direct Grant from CUHK, the National Key R&D Program of China (NO.2022ZD0160100), as well as cash and in-kind contribution from the industry partner(s).

## References

[1] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 2019. 2, 3, 8

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[3] Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Chen Change Loy, and Jinwei Gu. Removing diffraction image artifacts in under-display camera via dynamic skip connection networks. In *CVPR*, 2021. 1, 2, 6, 7

[4] Ruicheng Feng, Chongyi Li, Shangchen Zhou, Wenxiu Sun, Qingpeng Zhu, Jun Jiang, Qingyu Yang, Chen Change Loy, Jinwei Gu, et al. Mipi 2022 challenge on under-display camera image restoration: Methods and results. In *ECCVW*, 2022. 2

[5] KeMing Gao, Meng Chang, Kunjun Jiang, Yaxu Wang, Zhihai Xu, Huajun Feng, Qi Li, Zengxin Hu, and YueTing Chen. Image restoration for real-world under-display imaging. *Optics Express*, 29(23):37820–37834, 2021. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 6

[7] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2, 4

[8] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 7

[9] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *CVPRW*, 2020. 2, 3

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 5, 6

[11] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. TSIT: A simple and versatile framework for image-to-image translation. In *ECCV*, 2020. 7

[12] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *CVPR*, 2021. 3

[13] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Reference-based image and video super-resolution via $\mathcal{C}^2$-matching. *TPAMI*, 2022. 3

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2, 5

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3, 4

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3

[17] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *CVPR*, 2021. 6

[18] Irina Kim, Yunseok Choi, Hayoung Ko, Dongpan Lim, Youngil Seo, Jeongguk Lee, Geunyoung Lee, Eundoo Heo, Seongwook Song, and Sukhwan Lim. Under display camera quad bayer raw image restoration using deep learning. *Electronic Imaging*, 2021(7):67–1, 2021. 2

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[20] Jaihyun Koh, Jangho Lee, and Sungroh Yoon. BNUDC: A two-branched deep neural network for restoring images from under-display cameras. In *CVPR*, 2022. 7

[21] Kinam Kwon, Eunhee Kang, Sangwon Lee, Su-Jin Lee, Hyong-Euk Lee, ByungIn Yoo, and Jae-Joon Han. Controllable image restoration for under-display camera in smartphones. In *CVPR*, 2021. 2

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 4

[23] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 3

[24] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 6

[25] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *ECCV*, 2018. 2, 3, 5, 6

[26] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[27] Youngjin Oh, Gu Yong Park, Haesoo Chung, Sunwoo Cho, and Nam Ik Cho. Residual dilated u-net with spatially adaptive normalization for the restoration of under display camera images. In *APSIPA ASC*, 2021. 2

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[29] Miao Qi, Yuqi Li, and Wolfgang Heidrich. Isp-agnostic image reconstruction for under-display cameras. *arXiv preprint arXiv:2111.01511*, 2021. 1, 2

[30] Chengchao Qu, Ding Luo, Eduardo Monari, Tobias Schuchert, and Jürgen Beyerer. Capturing ground truth super-resolution data. In *ICIP*, 2016. 3

[31] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017. 8

[32] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, 2020. 2, 3

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 5

[34] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *CVPR*, 2019. 6

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[36] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 8

[37] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *CVPR*, 2021. 8

[38] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010. 5

[39] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 4, 5, 8

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[41] Andrea Vedaldi and Brian Fulkerson. VLFeat: An open and portable library of computer vision algorithms. In *ACMMM*, 2010. 3, 6

[42] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *ICCV*, 2021. 3

[43] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 5

[44] Xiangyu Xu and Chen Change Loy. 3d human texture estimation from a single image with transformers. In *ICCV*, 2021. 3

[45] Anqi Yang and Aswin C Sankaranarayanan. Designing display pixel layouts for under-panel cameras. *TPAMI*, 2021. 2

[46] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 3, 8

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[48] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *CVPR*, 2019. 2, 3, 6

[49] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *CVPR*, 2021. 3, 5

[50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 5

[51] Yuqian Zhou, Michael Kwan, Kyle Tolentino, Neil Emerton, Sehoon Lim, Tim Large, Lijiang Fu, Zhihong Pan, Baopu Li, Qirui Yang, et al. UDC 2020 challenge on image restoration of under-display camera: Methods and results. In *ECCVW*, 2020. 2

[52] Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. In *CVPR*, 2021. 1, 2